A Survey of Computer Vision Technologies In Urban and Controlled-environment Agriculture

JIAYUN LUO, Nanyang Technological University, Singapore

BOYANG LI*, Nanyang Technological University, Singapore

CYRIL LEUNG, Nanyang Technological University, Singapore and China-Singapore International Joint Research Institute, China

In the evolution of agriculture to its next stage, Agriculture 5.0, artificial intelligence will play a central role. Controlled-environment agriculture, or CEA, is a special form of urban and suburban agricultural practice that offers numerous economic, environmental, and social benefits, including shorter transportation routes to population centers, reduced environmental impact, and increased productivity. Due to its ability to control environmental factors, CEA couples well with computer vision (CV) in the adoption of real-time monitoring of the plant conditions and autonomous cultivation and harvesting. The objective of this paper is to familiarize CV researchers with agricultural applications and agricultural practitioners with the solutions offered by CV. We identify five major CV applications in CEA, analyze their requirements and motivation, and survey the state of the art as reflected in 68 technical papers using deep learning methods. In addition, we discuss five key subareas of computer vision and how they related to these CEA problems, as well as fourteen vision-based CEA datasets. We hope the survey will help researchers quickly gain a bird-eye view of the striving research area and will spark inspiration for new research and development.

CCS Concepts: • **Computing methodologies** → **Computer vision**; *Neural networks*; • **Applied computing** → **Agriculture**; • **General and reference** → *Surveys and overviews*.

Additional Key Words and Phrases: agriculture 5.0, controlled-environment agriculture, multimodality, pest and disease detection, growth monitoring, flower and fruit detection

1 INTRODUCTION

Artificial intelligence (AI), especially computer vision (CV), is finding an ever broadening range of applications in modern agriculture. The next stage of agricultural technological development, Agriculture 5.0 [15, 100, 236, 361], will constitute AI-driven autonomous decision making as a central component. The term Agriculture 5.0 stems from a chronology [361] that begins with Agriculture 1.0, which heavily depends on human labor and animal power, and Agriculture 2.0, enabled by synthetic fertilizers, pesticide, and combustion-powered machinery, and develops to Agriculture 3.0

*The authors can be reached at the following address: 50 Nanyang Avenue, School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. Boyang Li is the corresponding author. The research is funded by WeBank-NTU Joint Research Center and China-Singapore International Joint Research Institute.

and 4.0, characterized by GPS-enabled precision control, and Internet-of-Thing (IoT) driven data collection [257]. Built upon the rich agricultural data collected, Agriculture 5.0 holds the promise to further increase productivity, satiate the food demand of a growing global population, and mitigate the negative environmental impact of existing agricultural practices.

As an integral component of Agriculture 5.0, controlled-environment agriculture (CEA), a farming practice carried out within urban, indoor, resource-controlled, and sensor-driven factories, is particularly suitable for the application of AI and CV. This is because CEA provides ample infrastructure support for data collection and autonomous execution of algorithmic decisions. In terms of productivity, CEA could produce higher yield per unit area of land [8, 9] and boost the nutritional content of agricultural products [162, 313]. In terms of environmental impact, CEA farms can insulate environmental influences, relieve the need for fertilizer and pesticides, and efficiently utilize recycled resources like water, thereby may be much more environmentally friendly and self-sustainable than traditional farming.

In the light of current global challenges, such as disruptions to global supply chains and the threat of climate change, CEA appears especially appealing as a food source for urban population centers. Under pressures of deglobalization brought by geopolitical tensions [371] and global pandemics [237, 276], CEA provides the possibility to build farms close to large cities, which shortens the transportation distance and maintains secure food supplies even when long-distance routes are disrupted. The city-state Singapore, for example, has promised to source 30% of its food domestically by 2030 [1, 315], which is only possible through suburban farms such as CEAs. Furthermore, CEA, as a form of precision agriculture, is by itself a viable solution to the reduction of the emission of greenhouse gasses [9, 37, 249]. CEA can also shield plants from adverse climate conditions exacerbated by climate change as its environments are fully controlled [112] and is able to effectively reuse the arable land eroded due to climate change [373].

We argue that AI and CV are critical to the economic viability and long-term sustainability of CEAs as these technologies could save expenses associated with production and improve productivity. Suburban CEAs have high land costs. An analysis in Victoria, Australia [38] shows that, due to the higher land cost resulting from proximity to cities, with an estimated 50-fold productivity improvement per land area, it still takes 6 to 7 years for a CEA to reach the break-even point. Thus, further productivity improvement from AI would act as strong drivers for CEA adoption. Moreover, vertical or stacked setup of vertical farms impose additional difficulty for farmers to perform daily surveillance and operations. Automated solutions empowered by computer vision could effectively solve this problem. Finally, AI and CV technologies have the potential to fully characterize the complex, individually different, time-varying, and dynamic conditions of living organisms [39], which will enable precise and individualized management and further elevate yield. Thus, AI and CV technologies appear to be a natural fit to CEAs.

Most of the recent development of AI can be attributed to the newly discovered capability to train deep neural networks [175] that can (1) automatically learn multi-level representations of input data that are transferable to diverse downstream tasks [65, 137], (2) easily scale up to match the growing size of data [291], and (3) conveniently utilize massively parallel hardware architectures like GPUs [114, 337]. As function approximators, deep learning proves to be surprisingly effective in generalizing to previously unseen data [363]. Deep learning has achieved tremendous success in computer vision [302], natural language processing [47, 83, 118], multimedia [23, 88], robotics [300], game playing [278], and many other areas.

The AI revolution in agriculture is already underway. State-of-the-art neural network technologies, such as ResNet [134] and MobileNet [139] for image recognition, and Faster R-CNN [244], Mask R-CNN [133], and YOLO [239] for object detection, have been applied to the management of crops [197], livestock [142, 308], and plants in indoor and

vertical farms [245, 366]. AI has been used to provide decision support in a myriad of tasks from DNA analysis [197] and growth monitoring [245, 366] to disease detection [262] and profit prediction [28].

While several surveys have explored the use of computer vision (CV) techniques in agriculture, none of them specifically focus on CEA applications. Some surveys summarize studies based on aspects of practical applications in agriculture. [74, 89, 123, 149, 286] survey pest and disease detection studies. [40, 111, 312] discuss fruit and vegetable quality grading and disease detection. [307] summarizes studies in six sub-fields, including crop growth monitoring, pest and disease detection, automatic harvesting/fruit detection, fruit quality testing, automated management of modern farms and the monitoring of farmland information with Unmanned Aerial Vehicle (UAV). Other survey organize existing works from a technical perspective, namely algorithms used [241] or formats of data [56]. [154], as an exception, introduces the development history of CV and AI in smart agriculture, without investigating any individual studies. Our work aims to address this gap and provide insights tailored to CEA-specific contexts.
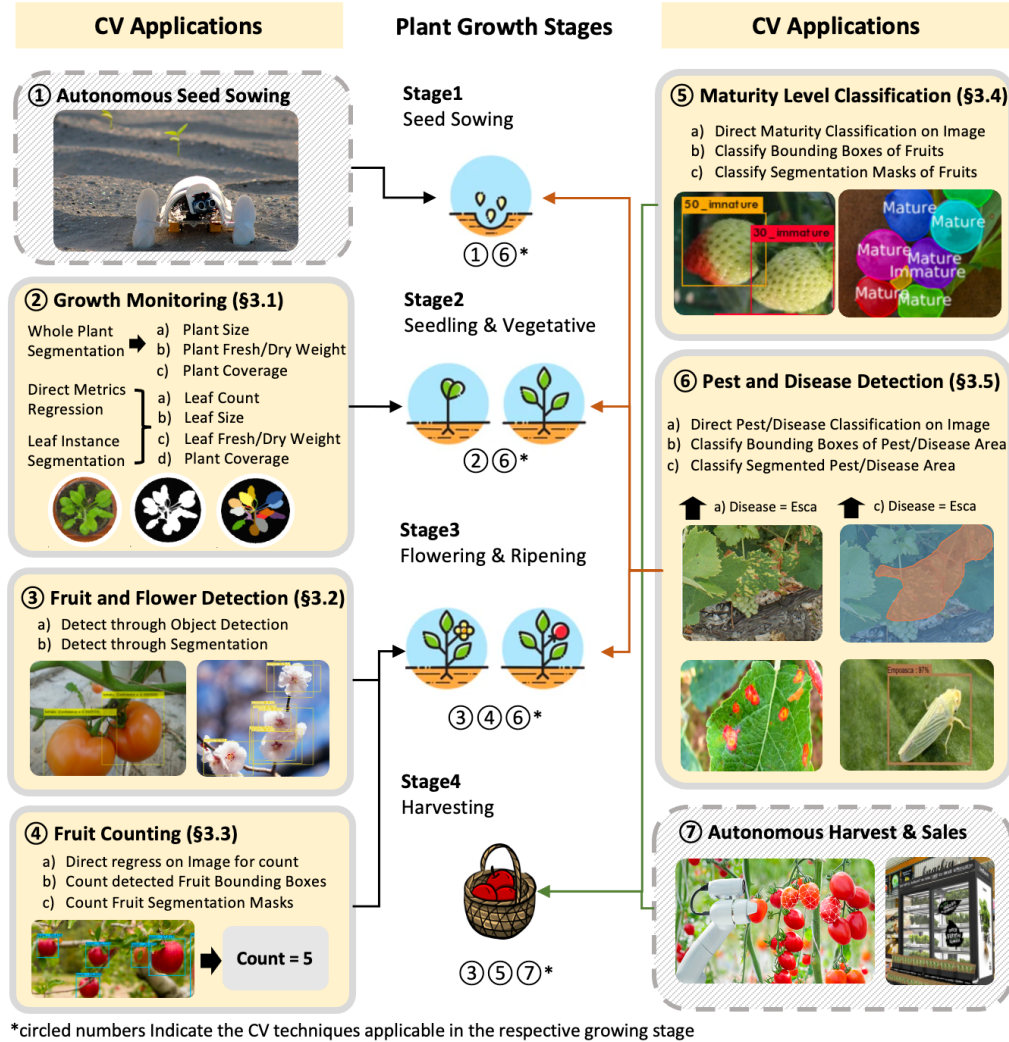
As the volume of research in smart agriculture grows rapidly, we hope the current review article can bridge researchers from AI and agriculture and create a mild learning curve when they wish to familiarize themselves in the other area. We believe computer vision has the closest connections with, and is the most immediately applicable in, urban agriculture and CEAs. Hence, in this paper, we focus on reviewing deep-learning based computer vision technologies in urban farming and CEAs. We focus on deep learning because it is the predominant approach in AI and CV research. The contributions of this paper are two-fold, with the former targeted at AI researchers and the latter targeted at agriculture researchers:

- We identify five major CV applications in CEA and analyze their requirements and motivation. Further, we survey the state of the art as reflected in 68 technical papers and 14 vision-based CEA datasets.
- We discuss five key subareas of computer vision and how they relate to CEA. In addition, we identify four potential future directions for research in CV for CEA.

In figure 1 we provide an graphical preview of our content. It illustrates the end-to-end agriculture process of CEAs, from seed planting to harvest and sales, with five major deep learning based CV applications–Growth Monitoring, Fruit and Flower Detection, Fruit Counting, Maturity Level Classification and Pest and Disease Detection–mapped to the corresponding applicable plant growth stages. We do not survey the autonomous seed planting and harvesting step as they are more relevant to robot functioning and robotic control, i.e grasping, carrying and placing of objects rather than computer vision (we do include the localization of fruit in the fruit and flower detection section that facilitate harvesting robot to locate the targeted object and perform action). However, we provide here some literature related to agriculture robot and end-effector design for reference [36, 57, 92, 235, 362]

We structure the survey following the process in the figure: First, to provide a bird-eye view of CV capabilities available to researchers in smart agriculture, we summarize several major CV problems and influential technical solutions in §2. Next, we review 68 papers with respect to the application of computer vision in the CEA system in §3. The discussion is organized into five subsections: Growth Monitoring, Fruit and Flower Detection, Fruit Counting, Maturity Level Classification, and Pest and Disease Detection. In the discussion, we focus on fruits and vegetables that are suitable for CEA, including tomato [10, 13, 127, 360], mango [7], guava [277, 342], strawberry [107, 355], capsicum [177], banana [5], lettuce [368], cucumber [10, 128, 203], citrus [4] and blueberry [2]. Next, we provide a summary of fourteen publicly available datasets of plants and fruits in §4 to facilitate future studies in Controlled-environment agriculture. Finally, we highlight a few research directions that could generate high-impact research in the near future in §5.

Fig. 1. An illustration of the end-to-end agriculture process of CEAs, from seed planting to harvest and sales, with five major deep learning based CV in agriculture applications–Growth Monitoring, Fruit and Flower Detection, Fruit Counting, Maturity Level Classification and Pest and Disease Detection – mapped to the corresponding applicable plant growth stages. Autonomous Seed Sowing and Autonomous Harvest and Sales in gray boxes are relevant steps in the agriculture process of CEAs but are out of the scope of our survey which focus on CV in CEAs. Orange lines represent arrows originated from pest and disease detection. Green lines represent arrows with stage 4 as destination.

CV Applications **Plant Growth Stages** **CV Applications**

① **Autonomous Seed Sowing**

Stage1
Seed Sowing

① ⑥*

⑤ **Maturity Level Classification (§3.4)**
a) Direct Maturity Classification on Image
b) Classify Bounding Boxes of Fruits
c) Classify Segmentation Masks of Fruits

② **Growth Monitoring (§3.1)**

Whole Plant Segmentation
a) Plant Size
b) Plant Fresh/Dry Weight
c) Plant Coverage

Direct Metrics Regression
a) Leaf Count
b) Leaf Size

Leaf Instance Segmentation
c) Leaf Fresh/Dry Weight
d) Plant Coverage

Stage2
Seedling & Vegetative

② ⑥*

⑥ **Pest and Disease Detection (§3.5)**
a) Direct Pest/Disease Classification on Image
b) Classify Bounding Boxes of Pest/Disease Area
c) Classify Segmented Pest/Disease Area

a) Disease = Esca   c) Disease = Esca

③ **Fruit and Flower Detection (§3.2)**
a) Detect through Object Detection
b) Detect through Segmentation

Stage3
Flowering & Ripening

③ ④ ⑥*

④ **Fruit Counting (§3.3)**
a) Direct regress on Image for count
b) Count detected Fruit Bounding Boxes
c) Count Fruit Segmentation Masks

Count = 5

Stage4
Harvesting

③ ⑤ ⑦*

⑦ **Autonomous Harvest & Sales**

*circled numbers Indicate the CV techniques applicable in the respective growing stage

One thing to note here is that, except for the Leaf Instance Segmentation task under the Growth Monitoring section, all the tasks are performed with model trained from different datasets and evaluated on different metrics. Table 3 4, 5, 6 showcase the variety in datasets and evaluation metrics. This variation results in incomparable performance between studies. Such a phenomenon further indicates the necessity of our survey, which summarizes the current progress in

literature and encourages the development of general benchmarks to promote consistency and comparability in future research.

2 COMPUTER VISION CAPABILITIES RELEVANT TO SMART AGRICULTURE

2.1 Image Recognition

The classic problem of image recognition is to classify an image containing a single object to the corresponding object class. The success of deep convolutional networks in this area dates (at least) back to LeNet [176] of 1998, which recognizes hand-written digits. The fundamental building block of such networks is the convolution operation. Using the principles of local connections and weight sharing, convolutional networks benefit from an inductive bias of translational invariance. That is, a convolutional network applies (approximately) the same operation to all pixel locations of the image.

The victory of AlexNet [166] in the 2012 ImageNet Large Scale Visual Recognition Challenge [253] is often considered as a landmark event that introduced deep neural networks into the AI mainstream. Subsequently, many variants of convolutional networks [151, 173, 279, 297] have been proposed. Due to space limits, here we provide a brief review of a few influential works, which is by no means exhaustive. ResNet [135] introduces residual connections that allow the training of networks of more than 100 layers. ResNeXT [345] and MobileNet [140] employ grouped convolution that reduces interaction between channels and improves the efficiency of the network parameters. ShuffleNet [374] utilizes the shuffling of channels, which complements group convolution. EfficientNet [301] shows simultaneous scaling of the network width, height, and image resolution is key to efficient use of parameters.

Recently, the transformer model has proven to be a highly competitive architecture for image recognition and other computer vision tasks [90]. These models cut the input image into a sequence of small image patches and often apply strong regularization such as RandAugment [75]. Variants such as CaiT [311], CeiT [359], Swin Transformer [198], and others [72, 78, 346, 380] achieve outstanding performance on ImageNet.

Despite the maturity of the technology for image classification, the assumption that an image contains only one object may not be easily satisfied in real-world scenarios. Thus, it is often necessary to adopt a problem formulation as object detection or semantic / instance segmentation.

2.2 Object Detection

The object detection task is to identify and locate all objects in the image. It can be understood as the task resulted from relaxing the assumption that the input image contains a single object. This is one natural problem formulation for real-world images and has seen wide adoption in agricultural applications.

In broad strokes, contemporary object detection methods can be categorized into anchor-box-based and point-based / proposal-free approaches. In anchor-box methods [110, 243], the process starts with a number of predefined anchor boxes that are periodically tiled to cover the entire input image. For each anchor box, the network makes two types of predictions. First, it determines if the anchor box contains one of the predefined object classes. Second, if the box contains an object, the network attempts to move and reshape the box to become closer to the ground-truth location of the object. One-stage anchor-box detectors [77, 101, 190, 196, 240, 376] make these predictions all at once. In comparison, two-stage detectors [110, 132, 189, 243], in the first stage discard anchor boxes that do not contain any object and classify the remaining boxes into finer object categories in the second stage. The location adjustment, known as bounding box regression, can happen in both stages. It is also possible to employ more than two stages [48]. When the objects have

diverse shapes and scales, these methods must create a large number of proposal boxes and evaluate them all, which can lead to high computational cost.

While point-based object detectors [91, 161, 174, 309, 381] still need to identify rectangular boxes around the objects, they make predictions at the level of grid locations on the feature maps. The networks predict if a grid location is a corner or the center of an object bounding box. After that, the algorithm assembles the corners and centers into bounding boxes. The point-based approaches can reduce the total number of decisions to be made. A careful comparison and analysis of anchor-box methods and point-based methods can be found in [369].

2.3 Semantic, Instance, and Panoptic Segmentation

Segmentation is a pixel-level classification task, aiming to classify every pixel in the image into a type of object or an object instance. The variations of the task differ by their definitions of the classes. In *semantic segmentation* [73, 94, 113, 170, 199], each type of object, such as cat, cow, grass, or sky, is its own class, but different instances of the same object type (*e.g.*, two cats) share the same class. In *instance segmentation* [76, 129, 131, 230], different instances of the same object type become unique classes, so that two cats are no longer the same class. However, object types such as sky or grass, which are not easily divided into instances, are ignored. In the recently proposed *panoptic segmentation* [69, 109, 158, 181, 192, 372], objects are first separated into things and stuff. Things are countable and each instance of things is its own class, whereas stuff is uncountable, impossible to separate into instances, appearing as texture or amorphous regions [12], and remains as one class. We note that the distinction between things and stuff is not rigid and can change depending on the application. For example, grass is typically considered as stuff, but in the leaf instance segmentation task, each leaf of a plant becomes an instance and is a separate class.

The primary requirement of pixel-level classification is to learn pixel-level representations that consider sufficient context and within reasonable computational budget. A typical solution is to introduce a series of downsampling followed by a series of upsampling operations. Since classic works such as the Fully Convolutional Network (FCN) [199] and U-Net [252], this has been the mainstream strategy for various segmentation strategies.

Due to its use in leaf segmentation, a problem in plant phenotyping, instance segmentation may be the most relevant segmentation formulation for urban farming. Despite the apparent similarity to semantic segmentation, instance segmentation poses challenges due to the variable number of instance classes and possible permutation of class indices [80]. This could be handled by combining proposal-based object detection and segmentation [61, 68, 129, 183, 231]. Mask-RCNN [132] exemplifies this approach. Leveraging its object detection capability, the network associates each object with a bounding box. After that, the network predicts a binary mask for the object within the bounding box. However, such methods may not perform well when there is substantial occlusion among objects or when objects are of irregular shapes [80].

Departing from the detect-then-segment paradigm, recurrent methods [242, 251, 258] that outputs one segmentation mask at one time may be considered as implicitly modeling occlusion. Pixel embedding methods [62, 80, 216, 225, 335, 340, 353] learn vector representations for every pixel and cluster the vectors. These methods are especially suitable for segmenting plant leaves and we will discuss them in greater detail in §3.1. Taking a page from the proposal-free object detector YOLO [239], SOLO [325] and SOLOv2 [326] divide the image into grids. The grid that the center an object falls into is responsible for predicting the segmentation mask of the object.

2.4   Uncertainty Quantification

Real-world applications often require qualification of the amount of uncertainty in the predictions made by machine learning, especially when the predictions carry serious implications. For example, if the system incorrectly determines that fruits are not mature enough, it may delay harvesting and cause overripe fruits with diminished values. Thus, users of the ML system are justified to ask how certain we are about the decision. In addition, when facing real-world input, it is desirable for the network to answer "I don't know" when facing an input that it does not recognize [186]. Well-calibrated uncertainty measurements may enable such a capability.

However, research shows that deep neural networks exhibit severe vulnerability to overconfidence, or under-estimation of the uncertainty in its own decisions [117, 206]. That is, the accuracy of the network decision is frequently lower than the probability that the network assigns to the decision. As a result, proper calibration of the networks should be a concern for systems built for real-world applications.

Calibration of deep neural networks may be performed post-doc (after training) using temperature scaling and histogram binning [87, 117, 321]. Also, regularization during training such as label smoothing [298] and mixup [138] have been shown to improve calibration [214, 228, 306]. Researchers propose new loss functions to replace existing ones that are susceptible to overconfidence [213, 352]. Moreover, ensemble methods such as Vertical Voting [344], Batch Ensemble [332], and Multi-input Multi-output [130] can derive uncertainty estimates.

2.5   Interpretability

Modern AI systems are known for its inability to provide faithful and human-understandable explanations for its own decisions. The unique characteristics of deep learning, such as network over-parameterization, large amount of training data, and stochastic optimization, while being beneficial to the predictive accuracy (e.g., [27, 182, 282, 289]), all create obstacles toward understand how and why a neural network reaches its decisions. The lack of human-understandable explanations leads to difficulties in the verification and trust of network decisions [52, 375].

We categorize model interpretation techniques into a few major classes, including visualization, feature attribution, instance attribution, inherently explainable models, and approximation by simple models. Visualization techniques present holistically what the model has learned from the training data by visualizing the model weights for direct visual inspection [34, 93, 97, 205, 212, 299]. In comparison, feature attribution and instance attribution are often considered as local explanations as they aim to explain model predictions on individual samples. Feature attribution methods [22, 58, 60, 210, 234, 263, 273, 281, 295, 349] generate a saliency map of an image or video frame, which highlights the pixels that contribute the most to its prediction. Instance attribution methods [32, 46, 67, 159, 233, 273, 350] attribute a network decision to training instances that, through the training process, exert positive or negative influence on the particular decision. Moreover, inherently explainable models [33, 59, 178, 267, 354] incorporate explainable components into the network architecture, which reduces the need to apply post-hoc interpretation techniques. In contrast, researchers also try to post-hoc approximate complex neural networks with simple models such as rule-based models [84, 102, 115, 156, 226, 327] or linear models [14, 105, 106, 163, 247] that are easily understandable.

The most significant benefit of interpretation in the context of CEA lies in its ability to aid with the auditing and debugging of AI systems and datasets. With feature attribution, users can make sure the system captures the robust features, or semantically meaningful features, that generalize to real-world data. As in the well-known case of husky vs. wolf image classification, due to a spurious correlation, the neural network learns to classify all images with white backgrounds as wolf and those with green backgrounds as husky [209]. Such shortcut learning can be identified by

feature attribution and subsequently corrected. Moreover, instance attribution allows researchers to pinpoint outliers or incorrectly labeled training data that may lead to misclassification [67].

3 CONTROLLED-ENVIRONMENT AGRICULTURE

Controlled-environment agriculture (CEA) is the farming practice carried out within urban, indoor, resource-controlled factories, often accompanied by stacked growth levels (*i.e.*, vertical farming), renewable energy and recycling of water and waste. CEA has recently been adopted in nations around the world [38, 82] such as Singapore [164], North America [6], Japan [9, 272], and UK [8].

CEA has economic and environmental benefits. Compared to traditional farming, CEA farms produce higher yield per unit area of land [8, 9]. Controlled environments shield the plants from seasonality and extreme weather, so that plants can grow all year round given suitable lighting, temperature and irrigation [38]. The growing conditions can as well be further optimized to boost growth and nutritional content [162, 313]. Rapid turnover increases farmers' flexibility in plant choice to catch the trend of consumption [35]. Moreover, farms investment on pesticides, herbicides, and transportation can be cut down due to reduced contamination from the outside environment and proximity to urban consumers.

CEA farms, when designed properly, can become much more environmentally friendly and self-sustainable than traditional farming. With optimized growing conditions and limited external interference, the need for fertilizer and pesticides decreases, so that we can reduce the amount of chemicals that go into the environment as well as the resulting pollution. Furthermore, CEA farms can save water and energy through the use of renewable energy and aggressive water recycling. For instance, CEA farms from Spread, a Japanese company, recycle 98% of used water and reduce the energy cost per head of lettuce by 30% with LED lightning [9]. Finally, CEA farm can be situated in urban or suburban areas, thereby reducing transportation and storage cost. A simulation for different farm designs in Lisbon shows vertical tomato farms with appropriate designs emit less greenhouse gas than conventional farms, mainly due to reduced water use and transportation distance [37].

A significant drawback of CEA, however, lies in its high cost, which may be partially addressed by computer vision technologies. According to [38], the higher land cost in Victoria, Australia means that the yield of vertical farms has to be at least 50 times more than traditional farming to break even. Computer vision holds the promise of boosting the level of automation and increasing yield, thereby making CEA farms economically viable. As would be discussed in the following sections, CV techniques can reduce a major amount of variable costs such as wastage cost induced by incorrect or delayed decisions on harvesting, and provide long-term benefit.

Carrying the potential to reduce a significant amount of cost, setting up computer vision systems in the field costs significantly less than expected when compared to the expenses of constructing a CEA building. Building a CEA structure involves high upfront costs, including construction, insulation, lighting, and HVAC systems. According to [272], a 1,300 square meter CEA building with a production area of 4,536 square meters would require a capital investment of $7.4 million and incur annual operational costs of approximately $3.4 million.

On the other hand, setting up hardware systems for CV models is relatively inexpensive. The necessary components include servers (CPU, GPU, memory, storage), sensors, cameras, networking, as well as cooling system. For example, a server with specifications like a 32-Core 2.80 GHz Intel Xeon Platinum 8462Y+, 128G memory, 4 NVIDIA RTX A6000 "Ada" GPUs, and 2TB storage costs around $60,000. Using this server for training purposes, assuming a standard VGG-16 architecture, training on 5000 images of size 224x224 pixels, with a batch size of 64 and 50 training epochs, and utilizing 4 NVIDIA A6000 GPUs, the estimated training time is less than an hour. Such a server is sufficient for daily training

Table 1. Factors to consider when applying CV techniques in CEA and some corresponding countermeasures.

Factors	CV Problems	Example Countermeasures
Environmental Change	OOD Generalization	Collect New Data, Few-shot learning, Weakly-supervised learning, Unsupervised-learning (see §3)
Sub-optimal Data Quality	Unbalanced Class Distribution, Lable Noise	Multiple-Instance Learning, Generate Image of Minority Classes with GANs, Few-shot Meta-learning (see §3.5.2, §3.5.3 and §5.1)
Human Factor	Interpretability, Uncertainty Estimates	Paired Confidence Scores, Meta-learning (see §2.4, §2.5 and §5.2)

and inference of commonly used CV models. For a camera system, if we consider 10 surveillance cameras such as the Hikvision DS-2CD2142FWD-I, the total cost would be around $1400. Additionally, a high-speed network infrastructure is required to transfer data between the computer hardware, storage, and camera systems. Typically it necessitates 4 to 7 routers to cover an area of 1300 square meters, costing approximately $2000. Finally, a liquid cooling system could cost between $1,000 and $2,000. In summary, a hardware system with a total cost of around $70,000 is sufficient for the daily operation, training, and inference of CV systems.

CEA can take diverse form factors [35] and the form factors may pose different requirements for computer vision technologies. Typical forms for CEA are glasshouses with transparent shells or completely enclosed facilities. Depending on the cultivars being planted, internal arrangement of the farm can be classified into stacked horizontal systems, vertical growth surfaces, and multi-floor towers. Form factors have influence on lighting, which is an important consideration in CV applications. For example, glasshouses with transparent shells utilize natural light to reduce energy consumption but may not provide sufficient lighting for CV around the clock. In comparison, a completely enclosed facility can have greater control of lighting conditions. Moreover, internal arrangement of the farm also affect camera angle. If cultivars being planted change frequently as a result of the high turnover rate in CEAs, the arrangement of shelves and plants might change. This would affect the camera angles and thus the resulting inference performance. CV systems need adapt to the change of the environment.

Nevertheless, with the autonomous setup of CEAs, which allow easy new data collection, training a new CV model or fine-tuning a previous model to adapt to the above mentioned changeable environment would be a cinch. Besides, there are also few-shot learning [294, 328], weakly-supervised learning[16, 222, 382] and unsupervised learning techniques [49, 261], which require minimal or zero annotations, that can facilitate the adjustment of the models.

Besides environmental change, there also exist other factors that need to be take into account when applying CV techniques in CEA. Two typical problems to consider would be 1) How to cope with sub-optimal data with label noise and how to address unbalanced class distribution. 2) How to interpret the prediction from models or measure the uncertainty of prediction so that users can use the models with confidence. Quantitative measure of the confidence or uncertainty would allow farmers to understand the decision generation process and make decisions with more confidence. Table 1 map these factors to consider into CV problems, and list corresponding solutions and the respective sections that discuss the solutions.

In the following, we investigate the application of autonomous computer vision techniques on Growth Monitoring, Fruit and Flower Detection, Fruit Counting, Maturity Level Classification and Pest and Disease Detection to increase

Table 2. Performance of various leaf instance segmentation techniques on the CVPPP A1 test set. Higher SBD and lower |DiC| indicate better performance. (GT-FG) indicates model making use of ground-truth foregrounds

Category	Technique	SBD (↑)	|DiC| (↓)
Sequential	End-to-end instance segmentation [242]	84.9	0.8
	RNN-SIS [259]	74.7	1.1
	RIS [251]	66.6	1.1
Pixel Embedding	Semantic Instance Segmentation [80]	84.2	1.0
	Object-aware Embedding [62]	83.1	**0.73**
	RHN + Cosine Embeddings [225]	84.5	1.5
	Crop Leaf and Plant Instance Segmentation [333]	91.1	1.8
	W-Net (GT-FG) [340]	91.9	-
	SPOCO (GT-FG) [336]	**93.2**	1.7

production efficiency. In addition to existing applications, we also include techniques that can be easily applied to vertical farms even though they have not yet been applied to them.

3.1 Growth Monitoring

Growth monitoring, a critical component of plant phenotyping, aims to understanding the life cycle of plants and estimating yield [148] by monitoring various growth indicators such as the plant size, number of leaves, leaf sizes, land area covered by the plant, and so on. Plant growth monitoring facilitates in quantifying the effects of biological / environmental factors on growth and thus is crucial for finding the optimal growing condition and developing high-yield crops [215, 303].

As early as 1903, Wilhelm Pfeffer has recognized the potential of image analysis in monitoring plant growth [229, 287]. Traditional machine vision techniques such as gray-level pixel thresholding [224], Bayesian statistics [45] and shallow learning techniques [150, 357], have been applied to segment the objects of interest, such as leaves and stems, from the background to analyze plant growth. Compared to traditional methods, deep-learning techniques provide automatic representation learning and are less sensitive to image quality variations. For this reason, deep learning techniques for growth monitoring have recently gained popularity.

Among various growth indicators, leaf size and number of leaves per plant are the most commonly used [121, 148, 172, 260]. Therefore, in the section below, we first discuss leaf instance segmentation, which can support both indicators at the same time, followed by a discussion of techniques for only leaf counting or for other growth indicators.

3.1.1 Leaf Instance Segmentation. Due to the popularity of the CVPPP dataset [207], the segmentation of leaf instance has attracted special attention from the computer vision community and warrants its own section. leaf instance segmentation methods include recurrent network methods [242, 251] and pixel embedding methods [62, 80, 225, 333, 340]. Parallel proposal methods are popular for general-purpose segmentation (see §segmentation), but are ill-suited for leaf segmentation. As most leaves have irregular shapes, the rectangle proposal boxes used in these methods do not fit the leaves well, resulting in many poorly positioned boxes. In addition, the density of leaves causes many proposal boxes to overlap and compounds the fitting problem. As a result, it is difficult to pick out the best proposal box from the large number of parallel proposals. Therefore, we focus on recurrent network based methods and pixel embedding based methods in this section. Quality metrics for leaf segmentation include Symmetric Best Dice (SBD) and Absolute

Difference in Count (|DiC|). SBD calculates the average overlap between the predicted mask and the ground truth for all leaves. DiC calculates the average number of miscalculated leaves over the entire test set.

Recurrent network based methods output a mask for a single leaf sequentially. Their decisions are usually informed by the already segmented parts of the image, which are summarized by the recurrent network. [242] applies LSTM and DeconvNet to segment one leaf at a time. The network first locates a bounding box for the next leaf, and performs segmentation within that box. After that, leaves segmented in all previous iterations are aggregated by the recurrent network and passed to the next iteration as contextual information. [251] employs convolution-based LSTMs (ConvL-STM) with FCN feature maps as input. At each time step, the network outputs a single-leaf mask and a confidence score. During inference, the segmentation stops when the confidence score drops below 0.5. [259] proposes another similar method that combines feature maps with different abstraction levels for prediction.

Pixel embedding methods learn vector representations for the pixels so that pixels in irregularly shaped leaves can become regularly shaped clusters in the representation space. With that, we can directly cluster the pixels. [333] performs simultaneous instance segmentation of leaves and plants. The authors propose an encoder-decoder framework, based on ERFNet [250], with two decoders. One decoder predicts the centroids of plants and leaves. The other decoder predicts the offset of each leaf pixels to the leaf centroid. The pixel location plus the offset vector hence should be very close to the leaf centroid. The dispersion among all pixels of the same leaf can be modeled as a Gaussian distribution, whose covariance matrix is also predicted by the second decoder and whose mean is from the first decoder. The training maximizes the Gaussian likelihood for all pixels of the same leaf. The same process is applied to pixels of the same plant.

[62, 225, 340] are three similar pixel embedding methods. They encourage pixels from the same leaf to have similar embeddings and pixels from different neighboring leaves to have different embeddings to enable clustering in the embedding space. Their network consists of two modules, the distance regression module and pixel embedding module. [225, 340] arrange the two modules in sequence, while [62] places them in parallel. The distance regression module predicts the distance between the pixel and the closest object boundary. The pixel embedding module generates an embedding vector for each pixel, so that pixels from the same leaves have similar embeddings and pixels from different neighboring leaves have different embeddings. During inference, pixels are clustered around leaf centers, which are identified as local maxima in the distance map from the distance regression module.

Lastly, [80, 336] take a large-margin approach. They ensure that embeddings of pixels from the same leaf are within a circular margin of the leaf center, and the embedding of leaf centers are far away from each other. This removes the need to determine the leaf centroids during inference because the embeddings are already well separated. [336] built upon the method in [80] to perform pixel embedding and clustering of leaves under weak supervision, with annotation on only a subset of instances in the images. In addition, a differentiable instance-level loss for a single leaf is formed to overcome the non-differentiability of assigning pixels to instances by comparing a Gaussian shape soft mask with the corresponding ground truth mask. Finally, consistency regularization, which encourages accordance of two embedding frameworks, is applied to improve embedding for unlabeled pixels.

Comparing different approaches, proposal-free pixel embedding techniques seem to be the best choice for the leaf segmentation problem. As can be seen from Table 2, pixel embedding methods obtain both the highest SBD and lowest |DiC|. One thing to note here, however, is that superior result of W-Net [340] and SPOCO [336] could be attributed to the inclusion of ground-truth foreground masks during inference. Even though the recurrent approach does not generate a large number of proposal boxes at once, it still uses rectangular proposals, which means that it still suffers from the fitting problem to irregular leaf shapes. Moreover, the recurrent methods are usually slower than pixel embeddings, due to the temporal dependence between the leaves.

Table 3. Performance of various fruit and flower detection techniques. Datasets without reference are unpublished datasets.

Category	Technique	Evaluation Metric	Performance	Dataset
Fruit Object Detection	[360]	Precision (IoU > 0.5)	94%	1730 images of cherry tomatoes
	[141]	Accuracy (IoU unspecified)	95.50%	800 images of tomatoes
	[255]	F1 scores (IoU unspecified)	83.80%	122 images of 7 fruits
	[365]	True positive rate and False positive rate (IoU unspecified)	98%, 17%	2116 self-acquired images of fruits and 511 images of fruits from ImageNet
	[355]	Precision and Recall (IoU > 0.9)	94.4%, 93.5%	2000 images of strawberries
	[270]	F1 scores (IoU unspecified)	93.5%-95.1%	Mango Image Dataset [160]
Fruit Segmentation	[188]	Precision and Recall (IoU unspecified)	98.3%, 94.8%	437 RGB-D images of guavas
	[13]	Precision, Recall and F1 scores (IoU > 0.5)	96%, 91%, 93%	123 images RGB-D images of tomatoes
	[107]	Precision, Recall, F1 score and Average Precision (IoU > 0.9)	97%, 92%, 94%, 90%	120 images RGB-D images of strawberries
	[356]	Mean IoU	89.85%	1900 images of strawberries
	[143]	Accuracy (IoU unspecified)	98%	900 images of strawberries
Flower Object Detection	[202]	Average Precision and F1 scores (IoU > 0.5)	96.2%, 89.0%	1078 images of citrus buds and flowers
	[292]	Average Precision (IoU > 0.5)	90.50%	5624 images of tomato flower and fruit
	[293]	IoU, F1 scores, Recall and Precision (IoU unspecified)	81.1%, 89.6%, 91.9%, 87.3%	Multi-species fruit flower detection [85]

3.1.2 Leaf Count and Other Growth Metrics. Leaf counts may be estimated without leaf segmentation. [314] utilizes synthetic data in the leaf counting task. The authors employ the L-system-based plant simulator *lpfg* [3, 232] to generate Arabidopsis rosette images. The authors test a CNN, trained with only synthetic data, on real data from CVPPP and obtain superior result than a model trained with CVPPP data only. In addition, CNN trained with the combination of synthetic and real data obtained approximately 27% reduction in the mean absolute count error compared to CNN using only real data. These results demonstrate the potential of synthetic data in plant phenotyping.

Besides leaf size and leaf count, leaf fresh weight, leaf dry weight, and plant coverage (the area of land covered by the plant) are also used as metrics of growth. [368] applies CNN to regress leaf fresh weight, leaf dry weight, and leaf area of lettuce on RGB images. [246] makes use of Mask R-CNN, a parallel proposal method, for lettuce instance segmentation. The authors derive plant attributes such as contour, side view area, height, and width from the segmentation masks and bounding boxes, using preset formulas. They also estimate growth rate from the changes in area of the plant at each time step; they estimate fresh weight by linearly regressing from the attributes. [201] leverages COCO dataset pretrained Mask R-CNN with ResNet-50 as backbone to segment lettuce leaves. The daily change of mean leaf area is used for growth rate calculation.

3.2 Fruit and Flower Detection

Algorithms for fruit and flower detection find the location and spatial distribution of fruits and fruit flowers. This task supports various downstream applications such as fruit count estimation, size estimation, weight estimation, robotic pruning, robotic harvesting, and disease detection [31, 108, 202, 351]. In addition, fruit or flower detection may help devise plantation management strategies [108, 127] because fruit or flower statistics such as positions, facing directions

(the directions the flowers face), and spatial scatter can reveal the status of the plant and the suitability of environmental conditions. For example, the knowledge of flower distribution may allow pruning strategies that focus on regions of excessive density and achieve even distribution of fruits which optimize the delivery of nutrient to the fruits.

Traditional approaches for fruit detection rely on manual feature engineering and feature fusion. As fruits tend to have unique colors and shapes, one natural thought is to apply thresholding on color [223, 331] and shape information [195, 221]. Additionally, [55, 187, 211] employ a combination of color, shape, and texture features. However, manual feature extraction suffers from brittleness when the image distribution changes with different camera resolutions, camera angles, illumination, and species [30].

Deep learning methods for fruit detection include object detection and segmentation. [360] applies SSD for cherry tomato detection. [141] leverages Faster R-CNN to detect tomatoes. Inside the generated bounding boxes, color thresholding and fuzzy-rule-based morphological processing methods are applied to remove image background and obtain the contours of individual tomatoes. [255] leverages Faster R-CNN with VGG-16 as the backbone for sweet pepper detection. RGB and near-infrared (NIR) images are used together for detection. Two fusion approaches, early and late fusion, are proposed. Early fusion alters the first pretrained layer to allow 4 input channels (RGB and NIR), whereas late fusion aggregates the two modalities by training independent proposal models for each modality and then combining the proposed boxes by averaging the predicted class probabilities. [365] trains three multi-task cascaded convolutional networks (MTCNN) [364] for detecting apples, strawberries and oranges. MTCNN contains a proposal network, a bounding box refinement network, and an output network in a feature pyramid architecture with gradually increased input sizes for each network. The model is trained on synthetic images, which are random combinations of cropped negative patches and fruits patches, in addition to real-world images. [355] proposed R-YOLO with MobileNet-V1 as the backbone to detect ripe strawberries. Different from regular horizontal bounding boxes in object detection, the model generates rotated bounding boxes by adding a rotation-angle parameter to the anchors.

Delicate fruits, such as strawberries and tomatoes, are particularly vulnerable to damage during harvesting. Therefore, much research has been devoted to segmenting such fruits from backgrounds in order to determine the precise picking point. Precise fruit masks are expected to enable robotic fruit picking while avoiding damages on the neighboring fruits. [188] performs semantic segmentation for guava fruits and determines their poses using FCN with RGB-D images as input. The FCN outputs a binary mask for fruits and another binary mask for branches. With the fruit binary mask, the authors employ Euclidean clustering [254] to cluster single guava fruit. From the clustering result and the branch binary mask, fruit centroids and the closest branch are located. Finally, the system predicts the vertical axis of the fruit as the direction perpendicular to the closest branch to facilitate robotic harvesting. Similarly, [13] leverages Mask R-CNN with ResNet as backbone for semantic segmentation of tomatoes. In addition, the authors filter the false positive detection of tomatoes from the non-targeted rows by setting a depth threshold. [107] utilizes Mask R-CNN with a ResNet101 backbone to perform instance segmentation of ripe strawberries, raw strawberries, straps and tables. Depth images are aligned with the segmentation mask to project the shape of strawberries into 3D space to facilitate automatic harvesting. [356] also applies Mask R-CNN with a ResNet101 + FPN backbone to perform instance segmentation and ripeness classification on strawberries. [143] leverages a similar network for instance segmentation of tomatoes. With the segmentation mask, the systems determine the cut points of the fruits.

Besides accuracy, the processing speed of neural networks is also important for their deployment on mobile devices or agricultural robots. [270] performs network pruning on YOLOv3-tiny to form a lightweight mango detection network. A YOLOv3-tiny pretrained on the COCO dataset has learned to extract fruit-relevant features because the COCO dataset contains apple and orange images, but it also has learned irrelevant features. The authors thus use a generalized

Table 4. Performance of various fruit counting techniques. Datasets without reference are unpublished datasets. [238] uses direct regression method thus does not need IoU threshold

Category	Technique	Evaluation Metric	Performance	Dataset
Count Regression	[238]	Accuracy	91.0% - 93%	4,800 synthetic tomato images
Count Fruit Bounding Boxes	[160] [329]	F1 scores, Average Precision(IoU > 0.24) R^2, RMSE	96.8%, 98.3% 0.66, 2.1	MangoYolo Dataset [160] MangoYolo Dataset [160]
Count Fruit Segmentation Masks	[157] [219]	Accuracy, F1 score (IoU > 0.6) Average Precision (IoU > 0.5), RMSE	73.6%, 84.4% 71.6%, 1.484	12,590 images of mangoes 724 images of blueberries

attribution method [274] to determine the contribution of each layer to fruit features extraction and remove convolution kernels responsible for detecting non-fruit classes. They find that the lower level features are shared across all classes detection and pruning in the higher layers does not harm fruit detection performance. After pruning, the network achieves significantly lowers float-point operations (FLOPs) at the same level of accuracy.

Object detection is also applied for flower detection. [202] proposes a modified YOLOv4-Tiny with cascade fusion (CFNet) to detect citrus buds, citrus flowers, and gray mold, which is a disease commonly found on citrus plants. The authors propose additionally a block module with channel shuffle and depth separable convolution for YOLOv4-Tiny. [292] shrinks the anchor boxes of Faster-RCNN to fit small fruits and applies soft non-maximum suppression to retain boxes that may contain occluded objects. As flowers usually have similar morphological characteristics, flowers from other non-targeted species could possibly be used as training data in a transfer learning scenario. In [293], the authors fine-tune a DeepLab-ResNet model [63] for fruit flower detection. The model is trained on apple flower dataset but achieves high F1 scores on pear and peach flower images (0.777 and 0.854 respectively).

3.3 Fruit Counting

Pre-harvest estimation of yields plays an important role in the planning of harvesting resources and marketing strategies [136, 347]. As fruits are usually sold to consumers as a pack of uniformly sized fruits or individual fruits, the fruit count also provides an effective yield metric [160], besides the distribution of fruit sizes. Traditional yield estimation is obtained through manual counting of samples from a few randomly selected areas [136]. Nonetheless, when the production is large-scale, to counteract the effect of plant variability, accurate estimation would require a large quantity of samples from different areas of the field, resulting in high cost. Thus, researchers resort to CV-based counting methods.

A direct counting method is to regress on the image and output the fruit count. In [238], the authors apply a modified version of Inception-ResNet for direct tomato counting. The authors train the model on simulated images and test on real images, which suggest, once again, the viability of using simulated images to circumvent the cost for formulating a large dataset.

Besides direct regression, object detection [160, 329], semantic segmentation [157], and instance segmentation [219] have also been used for fruit counting. These methods provide an intermediate level of results from which the count can be easily gathered. [160] proposes MangoYOLO based on YOLOv2-tiny and YOLOv3 for mango detection and counting. The authors increase the resolution of the feature map to facilitate detection of small fruits. [124] proposes pre-trained Faster R-CNN network, building upon DeepFruits [255], to estimate the quantity of sweet pepper. The authors design a tracking sub-system for sweet pepper counting. The sub-system identifies new fruits by measuring the IoU between and

14

Table 5. Performance of various maturity level classification techniques. Datasets without reference are unpublished datasets. Performance "-" are papers with unsummarizable metric results. [366] uses direct classification method thus does not need IoU threshold

Category	Technique	Evaluation Metric	Performance	Dataset
Classification	[366]	Accuracy	91.9%	200 images of tomatoes
Classification on Bounding Boxes	[355] [124]	Precision, Recall (IoU > 0.9) F1 score (IoU > 0.4)	94.4%, 93.5% 77.30%	2000 images of strawberries 285 images of capsicums
Classification on Segmentation Masks	[13]	Precision, Recall and F1 scores (IoU > 0.5)	-	123 images RGB-D images of tomatoes
	[107]	Precision, Recall, F1 score and Average Precision (IoU > 0.9)	-	120 images RGB-D images of strawberries
	[356]	Precision, Recall (IoU > 0.9)	95.78%, 95.41%	1900 images of strawberries
	[143]	Class frequency weighted precision and recall (IoU Unspecified)	96.1%, 96.0%	900 images of strawberries

comparing the boundary of detected and new fruits. [157] performs semantic segmentation for mango counting using a modification of FCN. The coordinates of blob-like regions in the semantic segmentation mask is used to generate bounding boxes corresponding to mango fruits. Finally, [219] applies Mask R-CNN to for instance segmentation of blueberries. The model also classifies the maturity of individual blueberries and counts the number of berries according to the masks.

Occlusion poses a difficult challenge for counting. Due to this issue, automatic count from detection or segmentation results is almost always lower than the actual number of fruits. To solve this, [160] calculates and applies the ratio between the actual hand harvest count and the automatic fruit count; it also uses both front and back views of mango trees to mitigate occlusion from one angle. Taking this idea one step further, [329] uses dual-view videos to detect and track mangoes when the camera moves. Utilizing different views of the same tree in a video, [329] recognizes around 20% more fruits. However, the detected count is still significantly lower than the actual number, underscoring the research challenge of exhaustive and accurate counting.

3.4 Maturity Level Classification

Maturity level classification aims to determine the ripeness of fruits or vegetables to aid in proper harvesting and food quality assurance. Premature harvesting results in plants that are unpalatable or incapable of ripening, while delayed harvesting can result in overripe plants or food decay [143].

The optimal maturity level differs for different targeted products and destinations. Fruits and vegetables can be consumed at different growing stages. For example, lettuce can be consumed either as baby lettuce or fully grown lettuce. The same situation happens with baby corn and normal corn. Products are to be transported to different destinations, so we must consider the length of transportation and ripening speed when deciding the correct maturity level at harvest [367].

Manually distinguishing the subtle differences in maturity levels is time-consuming, prone to inconsistency, and costly. The labor cost of harvesting accounts for a large percentage of operation cost in farms, with 42% of variable production expenses in U.S. fruit and vegetable farms being spent on labor for harvesting [144]. Automatic maturity level classification with computer vision, in contrast, can assist automatic harvesting [20, 107, 367] and reduce cost.

Similar to fruit detection, we can apply thresholding methods on color to detect ripeness. For example, [25] applies color thresholding on HSI and YIQ color spaces. [305] applies linear color models. [179] utilizes the combination of color and texture features. [96, 168, 264, 265, 338] apply shallow learning methods based on a multitude of features.

More recently, researchers evaluate the performance of deep learning based computer vision methods on maturity level classification and attain satisfactory results. For example, [366] applies CNN to classify tomato maturity into five levels. However, to further facilitate automatic harvesting, object detection and instance segmentation are more commonly used for getting the exact shape, location and maturity level of fruits, and position of peduncles for robotic end-effectors to cut on.

With object detection, [355] applies the R-YOLO network described in the fruit detection section (§3.2) to detect ripe strawberries. [124], as mentioned in the fruit counting section §3.3, proposes pre-trained Faster R-CNN network to estimate both the ripeness and quantity of sweet pepper. Two formulations of the model are tested. One treats ripe/unripe as additional classes on top of foreground/background, and the other performs foreground/background classification first and then performs ripeness classification on foreground regions. The second approach generates better ripeness classification results as the ripe/unripe classes are more balanced when only the foreground regions are considered.

Using the segmentation methods discussed in §3.2, [13] classifies semantic segmentation masks of tomatoes into raw and ripe tomatoes. [107, 356] performs instance segmentation and classifies instance masks into ripe and raw strawberries. [143] performs instance segmentation on tomatoes first. After transforming the mask region into HSV color space, the authors employ a fuzzy system to classify tomatoes into four classes: immature (completely green), breaker (green to tannish), preharvest (light red), and harvest (fully colored).

3.5 Pest and Disease Detection

Plants are susceptible to environmental disorders caused by temperature, humidity, nutritional excess/deficiency, light changes and biotic disorders due to fungi, bacteria, virus or other pests [103, 280]. Infectious diseases or pest pandemic induce inferior plant quality or plant death, resulting in at least 10% of global food production losses [290].

Although controlled vertical farming restricts the entry of pests and diseases, it cannot eliminate them. Pests and diseases can enter the farm from accidental contamination from employees, seeds, irrigation water and nutrient solution, poorly maintained environment or phytosanitation protocols, unsealed entrance and ventilation systems [248]. For this reason, pest and disease detection is still worth studying in the context of CEA.

Manual diagnosis of plant is complex due to the large quantity of vertically arranged plants in the field and numerous possible symptoms of diseases on different species. In addition, plants show different patterns along infection cycles and their symptoms can vary in different part of the plant [43]. Consequently, autonomous computer vision systems that recognize diseases according to the species and plant organs are gaining traction. From a technological perspective, we sort existing techniques into three parts, single- and multi-label classification, handling unbalanced class distributions, as well as label noise and uncertainty estimates.

3.5.1 Single- and Multi-label Classification. Studies perform single-label, or one-label-per-image, classification of diseases of either one single species [24, 262, 280, 370] or multiple species [95]. [370] creates a lightweight version of AlexNet, replacing the fully connected network with a global pooling layer, to classify six types of cucumber diseases. [280] leverages CNNs for classifying leaves into mango leaves, diseased mango leaves and other plant leaves. [24] utilizes

Table 6. Performance of various pest and disease detection techniques. Datasets without reference are unpublished datasets. Performance "-" are papers with unsummarizable metric results. *Studies perform direct classification on image thus do not need IoU threshold. [116] uses patch level segmentation which does not need IoU threshold as well.

Category	Technique	Evaluation Metric	Performance	Dataset
Single- and Multi-label Classification	[370]	Accuracy*	94.65%	700 diseased and normal leaf images
	[280]	Accuracy*	97.13%	1070 self acquired leaf images, and 1130 images from the Plant Village dataset [145]
	[24]	Accuracy*	93.33%	Images of 643 leaf samples
	[262]	mAP*	72.8% - 97.9%	12,600 images of bananas
	[95]	Accuracy*	99.50%	87,848 images of leaves
	[204]	Accuracy*	93.40%	Plant Village dataset [145]
	[103]	mAP (IoU > 0.5)	86%	5000 images of diseases and pests of tomatoes
	[358]	mIOU, recall, and F1-score (IoU unspecified)	84.8%, 88.1%, 91.8%	Plant Village dataset [145]
Handling Unbalanced Class Distribution	[44]	Accuracy*	60.7% - 91.8%	IP102 [339], Citrus Pest Benchmark [44]
	[116]	Average Precision	67% - 85%	Plant Village dataset [145] and Plant Leaves [275]
	[220]	Accuracy*	88.5% - 95.5%	Plant Village dataset [145] and Plant and Pest [185]
	[185]	Accuracy*	43.9% - 81%	Plant Village [145], Crop Pests Recognition [184]
Noise and Uncertainty Estimate	[271]	Accuracy*	94.58%	Plant Village [145]
	[99]	Accuracy*	-	15,892 images of tomatoes from Plant Village [145], extra 8911 images of corns, 6,635 images of soybeans

AlexNet and VGG16 to recognize five types of pests and diseases of tomatoes. [95] applies AlexNet, AlexNetOWTBn [165], GoogLeNet, Overfeat [266], and VGG for classifying 25 different healthy or diseased plants.

Having a single label per image can be inaccurate. In the real world, one plant or one leaf can carry multiple diseases or contain multiple diseased regions. By detecting multiple targeted areas or disease classes, the multi-label setting can lead to improved efficiency and accuracy.

To deal with the possibility of having multiple diseases or multiple areas of diseases on one plant simultaneously, two types of methods are proposed. [204] first segments out different infection areas on cucumber leaves using color thresholding following [203], then applies DCNN on segmented areas to classify four types of cucumber diseases. Nevertheless, the color thresholding technique may not generalize to other plant species and environment. Another type of method leverages object detection or segmentation for locating and classifying infection areas. [262] locates multiple diseased regions of banana plants simultaneously using object detection but assigns only one disease label to each image. [103] compared Faster R-CNN, R-FCN and SSD for detecting nine classes of diseases and pests that affect tomato plants. Multiple diseases and pests in one plant are detected simultaneously. [358] applies improved DeepLab v3+ for segmentation of multiple black rot spots on grape leaves. The efficient channel attention mechanism [324] is added to the backbone of DeepLab v3+ for capturing local cross-channel interaction. Feature pyramid network and Atrous Spatial Pyramid Pooling [64] are utilized for fusing feature maps from the backbone network at different scales to improve segmentation.

3.5.2 Handling Unbalanced Class Distributions. A common obstacle encountered in disease detection is unbalanced disease class distributions. There are typically much fewer diseased plants than healthy plants; the unequal frequencies introduce difficulties in finding images of rare diseases; the data unbalance leads to difficulty for model training. To

remedy such problem, researchers propose weakly supervised learning [44], generative adversarial network (GAN) [116], and few-shot learning [185, 220].

Specifically, [44] applies multiple instance learning (MIL), a type of weakly supervised learning method, for multi-class classification of six mite species of citrus. In MIL, the learner receives a set of labeled bags, containing multiple image instances. We know that at least one instance is associated with the class label, but do not know the exact instance. The MIL algorithm tries to identify the common characteristic shared by images in the positively labeled bags. In this work, a CNN is first trained with labeled bags. Next, by calculating saliency maps of images in bags, the model identifies salient patches that have a high probability of containing mites. These patches inherit labels from their bags and are used to refine the CNN trained above.

[116] leverages generative adversarial network (GAN) to generate realistic image patches of tip-burn lettuce and trains U-net for tip-burn segmentation. For the generation stage, lettuce canopy image patches are inputted into Wasserstein GANs [26] to generate stressed (tip-burned) patches so that there are an equal number of stressed and healthy patches. Then, in the segmentation stage, the authors generate a binary label map for the images using a classifier and an edge map. The binary label map labels each mini-patches (super-pixels) as stressed or healthy. The authors then feed the label map, alongside the original images, as input to U-net for mask segmentation.

In few-shot meta-learning, we are given a meta-train set and a meta-test set, with the two sets containing mutually exclusive image classes (i.e. classes in the training set do not appear in the testing set). Meta-train or meta-test sets contain a number of episodes, each of which consists of some training (supporting) images and some test (query) images. The rationale of meta-learning is to equip the model with the ability to quickly learn to classify the test images from a small number of training images within each episode. The model acquires this meta-learning capability on the meta-train set and is evaluated on the meta-test set.

As an example , [220] performs pests and diseases classification with few-shot meta-learning. The model framework consists of an embedding module and a distance module. The embedding module first projects supporting images into an embedding space using ResNet-18, then feeds embedding vectors into a transformer to incorporate information of other support samples in the same episode. After that, the distance module calculates the Mahalanobis distance[104] of the query and support samples to classify the query. Similarly, [185] uses a shallow CNN for embedding and the Euclidean distance for calculating the similarity between the embeddings of the query and support samples.

3.5.3 *Label Noise and Uncertainty Estimates.* [271] is another example of meta-learning, but it is used to improve the network's robustness against label noise. The model consists of two phrases. The first phrase is the conventional training of a CNN for classification. In the second phrase, the authors generate ten synthetic mini batches of images, containing real images with the labels taken from similar images. As a result, these mini-batches could contain noisy labels. After one step update on the synthetic instances, the network is trained to output similar predictions with the CNN from the first phrase. The result is a model that is not easily affected by noisy training data.

Finally, having a confidence score associated with the model prediction allows farmers to make decisions selectively under different confidence levels and boost the acceptance of deep learning models in agriculture. As an example, [99] performs classification of tomato diseases and pair the prediction with a confidence score following [79]. The confidence score, calculated using Bayes' rule, is defined as the probability of the true class label conditioned on the class probability predicted by the CNN. In addition, the authors build an ontology of disease classification. For example, the parent node "stressed plant" has as children "bacteria infection" and "virus infection", which in turn has "mosaic virus" as a child. If the confidence score of a specific terminal disease label is below a certain threshold, the model

switches to its more general parent label in the tree for higher confidence. By the axiom of probability, the predicted probability of the parent label is the summation of all the predicted probability of its direct descendants. For a general discussion of machine learning techniques that create well-calibrated uncertainty estimates, we refer readers to §2.4.

4 DATASETS

High-quality datasets with human annotations are one of the most important factors in the success of a machine learning project [208, 217, 334]. In this section, we review established datasets that enable training of CV models. We exclude datasets for plants that we have not found literature regarding their suitability in CEA, such as apples[41, 126], broccoli[169], and dates[21]. We have manually checked every dataset listed and assure that they are available for downloading at the time of writing. By summarizing the dataset related to CEA, we aim to facilitate interested researchers on their future studies. In the meantime, we would like to encourage scholars to publish more datasets dedicated to CEA.

As listed in Table 7 and Table 8, we discover fourteen datasets in CEA, with three for Growth Monitoring, five for Fruit Detection, and six for Pest and Disease Detection. Each targeted task contains at least one dataset that covers multiple species to facilitate training of generalizable and transferable models. The largest dataset is CVPPP with 6,287 and 165,120 RGB images for Arabidopsis and Tobacco respectively, aiming for growth monitoring related tasks. All the available datasets are composed of real images. While real images provide realistic data, we also want to encourage publication of synthetic datasets, which usually feature balanced class distribution and accurate labeling. Another point noteworthy is that many real images are collected under simplified laboratory environments, which may bias the data toward specific lighting conditions, backgrounds, plant orientation, or camera positions. For real world application, practitioners may need to further finetune the trained models on more realistic data.

5 FUTURE RESEARCH DIRECTIONS

So far we have discussed the objectives, benefits, and realizations of Growth Monitoring, Fruit and Flower Detection, Fruit Counting, Maturity Level Classification, and Pest and Disease Detection in CEA precision farming. Based on the current research status and existing technical capabilities of computer vision, we would like to point out several areas where computer vision technologies could provide short- to mid-term benefits to urban and suburban CEA. We identify three such areas, including realistic datasets that are unbalanced and noisy, uncertainty quantification, and multi-task learning / system integration.

5.1 Handling Realistic Data

The ability to handle realistic data is a critical competence that has not received sufficient research attention (with a few notable exceptions [44, 116, 185, 220, 271]). Unlike well-curated datasets that have accurate and abundant labels and relatively balanced label distributions, real-world data exhibit skewed label distribution as well as substantial noise in the labels. For effective real-world application, it is important that the CV algorithms can maintain good predictive performance under these conditions. In addition, the algorithmic tolerance of data imperfection can lower annotation cost and enable wider applications of CV. There has been substantial research on these topics in the computer vision community, such as long-tail recognition [81, 193, 269, 322, 378, 379], few-shot and zero-shot learning [180, 283–285, 343], as well as noise-resistant classification [17, 70, 153, 330, 377] and metric learning [147, 191, 319]. We believe that research on smart agriculture could benefit from the existing body of literature.

Table 7. Dataset for CV tasks in CEA

Target Task	Dataset	Release Year	Data Description	URL
Growth Monitoring	CVPPP dataset [207]	2014	6,287 and 165,120 RGB images (resolution 72x72) of Arabidopsis and Tobacco respectively . Annotations include bounding boxes and segmentation masks for every plant and every leaf, and the leaf centers.	https://www.plant-phenotyping.org/datasets-download
	Oil Radish dataset [167]	2019	129 RGB images (resolution 1x1) of oil radish with binary semantic segmentation mask and respective plant fresh and dry weight, as well as nutrient content.	https://competitions.codalab.org/competitions/20981#learn_the_details
	Leaf Counting dataset [304]	2018	9,372 RGB images (resolution 72x72) of weeds with the number of leaves counted.	https://vision.eng.au.dk/leaf-counting-dataset/
Fruit and Flower Detection	DeepFruits [256]	2016	RGB images (resolution 72x72 to 400x400) of sweet pepper, rock melon, apple, mango, orange and strawberry images annotated with rectangular bounding boxes. Each fruit has 42-170 images.	https://drive.google.com/drive/folders/1CmsZb1caggLRN7ANfika8WuPiywo
	Orchard Fruit[30]	2016	1,120, 1,964 and 620 RGB images (resolution 72x72) of apple, mango and almond, respectively. Apples are annotated with bounding circles; mango and almond are annotated with rectangular bounding boxes	http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/
	MangoYOLO [160]	2019	1,730 RGB images of mango (resolution 72x72 and 300x300), annotated with rectangular bounding boxes; photos are under artificial lighting	https://figshare.com/articles/dataset/MangoYOLO_data_set/13450661
	MangoNet Semantic Dataset[157]	2019	45 training images and 4 test images (resolution 180x180) of mango. Each image is annotated with semantic segmentation mask which is colored green in regions of mangoes and black in non mango regions.	https://github.com/avadesh02/MangoNet-Semantic-Dataset
	Fruit Flower detection [86]	2018	162, 20, and 15 images (resolution 72x72) of apple, peach, and pear flowers annotated with binary semantic segmentation mask with white represents flower pixels.	https://data.nal.usda.gov/dataset/data-multi-species-fruit-flower-detection-using-refined-semantic-segmentation-network

5.2 Quantifying Uncertainty and Interpretability

Real-world applications call for reliable estimation of the quality of automated decisions. An incorrect prediction made by an AI system may have profound implications. For example, if the system incorrectly determines that fruits are not mature enough, it may delay harvesting and cause overripe fruits with diminished values. However, it is impossible to eliminate incorrect or uncertain predictions, as they originate from factors difficult to control and precisely measure, including model assumptions, test data shift, incomplete training data and so on [11, 146]. Thus, we argue that uncertainty quantification is another crucial factor for real-world deployment. Such quantification would allow

Table 8. Dataset for CV tasks in CEA

Target Task	Dataset	Release Year	Data Description	URL
Pest and Disease Detection	Plant Village [145]	2019	61,486 RGB images (resolution 72x72) of plant leaves, with 39 different classes of diseased and healthy plant leaves	https://data.mendeley.com/datasets/tywbtsjrjv/1
	Crop Pests Recognition [184]	2020	5,629 RGB images (resolution 72x72) of 10 pest classes, each class containing over 400 images	https://bit.ly/2DdUFza
	Plant and Pest [185]	2021	6,000 RGB images (resolution 72x72) of 20 different classes of plant leaves and pests from Plant Village [145] and Crop Pests Recognition [184]	https://zenodo.org/record/4529076#.YupE_-xBzlw
	Citrus Pest Benchmark [44]	2022	10,816 multi-class RGB images (resolution 1200x1200) categorized into seven classes of pests	https://github.com/edsonbollis/Citrus-Pest-Benchmark
	IP102 [339]	2019	75,000 images (resolution 400x300) of 102 insect classes and among these 19,000 are annotated with bounding boxes.	https://github.com/xpwu95/IP102
	Plant Leaves [275]	2022	4503 images (resolution 6000x4000) of which contains 2278 images of healthy leaves and 2225 images of the diseased leaves	https://data.mendeley.com/datasets/hb74ynkjcn/1

farmers to make informed decisions on whether to follow the machine recommendation or not. For the convenience of readers, we provide a brief review of such deep learning techniques in §2.4.

Besides uncertainty quantification, pair the model with explanation on its decisions could enhance user confidence and assist auditing and debugging of the AI system. Specifically, instance attribution methods, as discussed in §2.5, enable detection of the biased or low quality data points with extreme influence on prediction [67]. For example, if the model is trained with an image of dry leaves with dust that resemble a certain disease of the plant, in the inference process, the model might misclassify diseased leaves as normal dry leaves or vice versa and induce plant death or unnecessary treatments. With instance attribution interpretation, researchers can identify misleading data points and perform adversarial training to improve model accuracy.

5.3 Multi-task Learning and System Integration

Real-world deployment usually requires the coordination of multiple CV capabilities provided by different networks. When the system is designed well, these networks could facilitate each other and achieve synergistic effects. For example, instance segmentation can be used for fruit and flower localization (§3.2), growth monitoring (§3.1), and fruit maturity level detection (§3.4). However, academic research tends to study these problems in isolation, thereby unable to reap benefits of multi-task learning.

Multi-task learning [29, 51, 194] focuses on leveraging mutually beneficial supervisory signals from multiple correlated tasks. Recently, CV researchers have built large-scale networks [66, 71, 120, 152, 155, 200, 323, 383] that perform a wide

range of tasks and achieve state-of-the-art results on most tasks. This demonstrates the benefits of multi-task learning and could inspire similar work dedicated to smart farming in CEAs.

Another motivation for considering multi-task learning and system integration is that errors can propagate in a pipeline architecture. For example, a network could first incorrectly detect a leaf occluding a mature fruit as the fruit and then classify it as an immature fruit. As a result, simply concatenating multiple techniques will result in inferior overall performance than what practitioners may expect. Thus, we encourage system designers to consider end-to-end training, or other innovative techniques [119, 341, 348] for aligning and interfacing different components within a system.

Finally, multi-task learning handles multiple tasks simultaneously, which saves computation power, enhances data efficiency, and alleviates the necessity to maintain and iterate multiple models. Such benefits are crucial for popularizing CEAs, as they facilitate the efficient use of energy, computation power, and human resources. Consequently, both the initial setup and ongoing maintenance investments for CEA farms can be reduced, expediting the emergence of economically viable CEAs. Furthermore, mindful selection and combination of targeted tasks have the potential to further improve overall efficiency [288].

5.4 Effective Use of Multimodality

Fusion of multi-modal data enhances inference ability of models by incorporating complementary view of data [171]. In the context of CEA, thermal or depth images capture the depth or temperature differences between foreground and background and enable filtering of non-target objects (e.g., fruits or leaves). Abnormal temperature changes during growth cycle can also indicate disease infection before visual symptoms appear [53, 54]. Furthermore, as different materials absorb, reflect, and transmit light in different ways and at different wavelengths, multi-spectral imaging (MSI) and hyper-spectral imaging (HSI), which capturing images at multiple wavelengths of light, can be used to perform more specific internal inspection of leaves, fruits and plants as compared to thermal and depth images. Finally, LiDAR and RGB-D systems allow the generation of high density 3D point clouds of plants, fruits [107, 188] or environment [318], which facilitate 3D volume measurement or cut-point detection during harvesting.

Existing works have demonstrated the efficacy of multi-spectral imaging (MSI) and hyper-spectral imaging (HSI) [13, 42, 320]. MSI have been utilized for yield prediction [310] and early disease detection [227, 316]. However, current literature explored majorly the power of MSI with shallow machine learning. We found only one work that leverages deep learning on MSI input [310], which applies a pruned VGG-16 for wheat yield estimation. HSI provides finer-grained resolution and divides the range of wavelength into many more spectral bands than MSI, typically ranging from tens to hundreds of bands, though at a higher cost. Hyper-spectral images have been used as the sole modality in early disease detection with both shallow machine learning methods [18, 19, 296] and deep learning methods [98, 122, 218, 320]. Due to relevancy and space limit, we will only talk about the deep learning methods here. Specifically, with a GAN-based data augmentation method, [320] performs early detection of tomato spotted wilt virus before visible symptoms appear using hyper-spectral images. [218] performs early detection of grapevine vein-clearing virus and shows the discriminative power of HSI in combination with CNN and shallow machine learning algorithms. [98] attains early barley disease detection through generating future prediction of hyper-spectral barley leaf images using GAN. Moreover, HSI has also been utilized for yield prediction through fruit counting. [122] leverages CNN and HSI to segment semantic mango masks and count the number of fruits.

However, systematic exploration of fusion techniques for multimodal inputs remains relatively rare in CEA applications. Many existing approaches adopt pipeline-based multimodal integration techniques that do not exhaust the

potential of deep learning due to the lack of end-to-end training. For example, in [13], the authors set a depth threshold to filter false positive tomato detection from the background. [42] first performs broccoli segmentation on the RGB image. Within the segmentation mask, the authors find the mode of the depth value distribution, which is used to calculate the diameter of the broccoli head. [188] conducts semantic segmentation for guava fruits using RGB images and reconstructs their 3D positions from the depth input. [107] utilizes Mask R-CNN to perform instance segmentation of strawberries and align depth image with the segmentation mask to obtain 3D shape of strawberries. These methods use the two modalities separately and do not apply end-to-end training of the pipeline. As exceptions, [255] proposes late fusion of RGB and near-infrared images in sweet pepper detection. [317] incorporates depth information by replacing the blue channel with depth channel and applies masked R-CNN to locate tomatoes.

In computer vision research, numerous techniques for fusing and joint utilization of multimodal information have been proposed over the years, which we believe could contribute to CV applications in CEA. Due to space limits, we list only a few examples here. [268] proposes two different ways to combine multiple modalities in object detection, Concatenation and Element-wise Cross Product. The former combines feature maps from different modalities along the channel dimension and let the network discover the best way to combine them from data. The latter technique, Element-wise Cross Product, applies element-wise multiplication to every possible pair of feature maps from the two modalities. [50] experiments with a variety of fusion techniques for RGB and optical flow and discovers a high-performing late-fusion strategy in action recognition. In self-supervised learning, [125] identifies similar data points using one modality and treats them as positive pairs in another modality. This technique provides another paradigm to leverage the complementary nature of multimodality.

6   CONCLUSIONS

Smart agriculture, and particularly computer vision for controlled-environment agriculture (CV4CEA), are rapidly emerging as an interdisciplinary area of research that could potentially lead to enormous economic, environmental and social benefits. In this survey, we first provide brief overviews of existing CV technologies that range from image recognition to structured understanding such as segmentation; from uncertain quantification to interpretable machine learning. Next, we systematically review existing applications of CV4CEA, including growth monitoring, fruit and flower detection, fruit counting, maturity level classification, and pest / disease detection. Finally, we highlight a few research directions that could generate high-impact research in the near future.

Like any interdisciplinary area, research progress in CV4CEA requires expertise in both computer vision and agriculture. However, it could take a substantial amount of time for any researcher to acquire in-depth understanding of both subjects. By reviewing existing applications, available CV technologies, and identifying possible future research directions, we aim to provide a quick introduction of CV4CEA to researchers with expertise in agriculture or computer vision alone. It is our hope that the current survey will serve as a bridge between researchers from diverse backgrounds and contribute to accelerated innovation in the next decade.

7   ACKNOWLEDGMENTS

REFERENCES

[1] [n. d.]. 30 by 30: Strengthening our food security. https://www.ourfoodfuture.gov.sg/30by30. Accessed: 2022-8-15.

[2] [n. d.]. AeroFarms Partners With Hortifrut to Grow Blueberries, Caneberries Via Vertical Farming. https://thespoon.tech/aerofarms-partners-with-hortifrut-to-grow-blueberries-caneberries-via-vertical-farming/. Accessed: 2022-7-28.

[3] [n. d.]. Algorithmic Botany. http://www.algorithmicbotany.org/virtual_laboratory/. Accessed: 2022-6-20.

[4] [n. d.]. All in(doors) on citrus production. https://www.hortibiz.com/newsitem/news/all-indoors-on-citrus-production/. Accessed: 2022-7-28.

[5] [n. d.]. Greenhouse in Shanghai successfully plants bananas on water. https://www.hortidaily.com/article/9369964/greenhouse-in-shanghai-successfully-plants-bananas-on-water/. Accessed: 2022-7-28.

[6] [n. d.]. Introducing VertiCrop™. https://verticrop.com/. Accessed: 2022-5-24.

[7] [n. d.]. Mango trees cultivation under greenhouse conditions. https://horti-generation.com/mango-trees-cultivation-under-greenhouse-conditions/. Accessed: 2022-7-28.

[8] [n. d.]. Saturn Bioponics. http://www.saturnbioponics.com/. Accessed: 2022-05-25.

[9] [n. d.]. Spread-A new way to grow vegetable. https://spread.co.jp/en/environment/. Accessed: 2022-05-24.

[10] [n. d.]. Tomatoes and cucumbers in a vertical farm without daylight. https://www.hortidaily.com/article/9212847/tomatoes-and-cucumbers-in-a-vertical-farm-without-daylight/. Accessed: 2022-7-28.

[11] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297.

[12] Edward H. Adelson. 2001. On seeing stuff: the perception of materials by humans and machines. In *Human Vision and Electronic Imaging VI*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas (Eds.), Vol. 4299. International Society for Optics and Photonics, SPIE, 1 – 12. https://doi.org/10.1117/12.429489

[13] Manya Afonso, Hubert Fonteijn, Felipe Schadeck Fiorentin, Dick Lensink, Marcel Mooij, Nanne Faber, Gerrit Polder, and Ron Wehrens. 2020. Tomato fruit detection and counting in greenhouses using deep learning. *Frontiers in plant science* 11 (2020), 571299.

[14] Isaac Ahern, Adam Noack, Luis Guzman-Nateras, Dejing Dou, Boyang Li, and Jun Huan. [n. d.]. NormLime: A New Feature Importance Metric for Explaining Deep Neural Networks. *arXiv Preprint 1909.04200* ([n. d.]). https://doi.org/10.48550/ARXIV.1909.04200

[15] Latief Ahmad and Firasath Nabi. 2021. *Agriculture 5.0: Artificial Intelligence, IoT and Machine Learning.* CRC Press.

[16] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2209–2218.

[17] Görkem Algan and Ilkay Ulusoy. 2020. Meta soft label generation for noisy labels. In *ICPR*.

[18] Ali AlSuwaidi, Bruce Grieve, and Hujun Yin. 2018. Combining spectral and texture features in hyperspectral image analysis for plant monitoring. *Measurement Science and Technology* 29, 10 (2018), 104001.

[19] Ali AlSuwaidi, Bruce Grieve, and Hujun Yin. 2018. Feature-ensemble-based novelty detection for analyzing plant hyperspectral datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 4 (2018), 1041–1055.

[20] H Altaheri, M Alsulaiman, M Faisal, and G Muhammed. 2019. Date fruit dataset for automated harvesting and visual yield estimation. In *Proc. IEEE DataPort*.

[21] Hamdi Altaheri, Mansour Alsulaiman, Mohammed Faisal, and Ghulam Muhammed. 2019. Date Fruit Dataset for Automated Harvesting and Visual Yield Estimation. https://doi.org/10.21227/x46j-sk98

[22] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104* (2017).

[23] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Krishnaswamy R Aravind, Purushothaman Raja, Rajendran Ashiwin, and Konnaiyar V Mukesh. 2019. Disease classification in Solanum melongena using deep learning. *Spanish Journal of Agricultural Research* 17, 3 (2019), e0204–e0204.

[25] Arman Arefi, Asad Modarres Motlagh, Kaveh Mollazade, and Rahman Farrokhi Teimourlou. 2011. Recognition and localization of ripen tomato based on machine vision. *Australian Journal of Crop Science* 5, 10 (2011), 1144–1149.

[26] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[27] Sanjeev Arora, Nadav Cohen, and Elad Hazan. 2018. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*. PMLR, 244–253.

[28] S Aruul Mozhi Varman, Arvind Ram Baskaran, S Aravindh, and E Prabhu. 2017. Deep Learning and IoT for Smart Agriculture Using WSN. In *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. 1–6. https://doi.org/10.1109/ICCIC.2017.8524140

[29] BJ Bakker and TM Heskes. 2003. Task clustering and gating for bayesian multitask learning. (2003).

[30] Suchet Bargoti and James Underwood. 2017. Deep fruit detection in orchards. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3626–3633.

[31] Suchet Bargoti and James P Underwood. 2017. Image segmentation for fruit detection and yield estimation in apple orchards. *Journal of Field Robotics* 34, 6 (2017), 1039–1060.

[32] Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1899–1909.

[33] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160* (2019).

[34] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[35] Andrew M Beacham, Laura H Vickers, and James M Monaghan. 2019. Vertical farming: a summary of approaches to growing skywards. *The Journal of Horticultural Science and Biotechnology* 94, 3 (2019), 277–283.

[36] Avital Bechar and Clément Vigneault. 2016. Agricultural robots for field operations: Concepts and components. *Biosystems Engineering* 149 (2016), 94–111.

[37]   Khadija Benis, Christoph Reinhart, and Paulo Ferrão. 2017. Development of a simulation-based decision support workflow for the implementation of Building-Integrated Agriculture (BIA) in urban contexts. *Journal of cleaner production* 147 (2017), 589–602.

[38]   Kurt Benke and Bruce Tomkins. 2017. Future food-production systems: vertical farming and controlled-environment agriculture. *Sustainability: Science, Practice and Policy* 13, 1 (2017), 13–26.

[39]   Daniel Berckmans. 2017. General introduction to precision livestock farming. *Animal Frontiers* 7, 1 (2017), 6–11.

[40]   Anuja Bhargava and Atul Bansal. 2021. Fruits and vegetables quality evaluation using computer vision: A review. *Journal of King Saud University-Computer and Information Sciences* 33, 3 (2021), 243–257.

[41]   Santosh Bhusal, Manoj Karkee, and Qin Zhang. 2019. Apple Dataset Benchmark from Orchard Environment in Modern Fruiting Wall. (2019).

[42]   Pieter M. Blok, Eldert J. van Henten, Frits K. van Evert, and Gert Kootstra. 2021. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosystems Engineering* 208 (2021), 213–233. https://doi.org/10.1016/j.biosystemseng.2021.06.001

[43]   CH Bock, PE Parker, AZ Cook, and TR Gottwald. 2008. Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. *Plant Disease* 92, 4 (2008), 530–541.

[44]   Edson Bollis, Helio Pedrini, and Sandra Avila. 2020. Weakly supervised learning guided by activation mapping applied to a novel citrus pest benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 70–71.

[45]   Charles A Bouman and Michael Shapiro. 1994. A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on image processing* 3, 2 (1994), 162–177.

[46]   Jonathan Brophy and Daniel Lowd. 2020. TREX: Tree-Ensemble Representer-Point Explanations. *arXiv preprint arXiv:2009.05530* (2020).

[47]   Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv 2005.14165* (2020).

[48]   Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49]   Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[50]   Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[51]   Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[52]   Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[53]   Laury Chaerle, Dik Hagenbeek, Erik De Bruyne, Roland Valcke, and Dominique Van Der Straeten. 2004. Thermal and chlorophyll-fluorescence imaging distinguish plant-pathogen interactions at an early stage. *Plant and Cell Physiology* 45, 7 (2004), 887–896.

[54]   Laury Chaerle, Ilkka Leinonen, Hamlyn G Jones, and Dominique Van Der Straeten. 2007. Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. *Journal of experimental botany* 58, 4 (2007), 773–784.

[55]   Supawadee Chaivivatrakul, Jednipat Moonrinta, and Matthew N Dailey. 2010. Towards Automated Crop Yield Estimation-Detection and 3D Reconstruction of Pineapples in Video Sequences.. In *VISAPP (1)*. 180–183.

[56]   Akshay L Chandra, Sai Vikas Desai, Wei Guo, and Vineeth N Balasubramanian. 2020. Computer vision with deep learning for plant phenotyping in agriculture: A survey. *arXiv preprint arXiv:2006.11391* (2020).

[57]   Fernando Alfredo Auat Cheein and Ricardo Carelli. 2013. Agricultural robotics: Unmanned robotic service units in agricultural tasks. *IEEE industrial electronics magazine* 7, 3 (2013), 48–58.

[58]   Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 782–791.

[59]   Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can Rationalization Improve Robustness? *arXiv preprint arXiv:2204.11790* (2022).

[60]   Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015* (2020).

[61]   Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4974–4983.

[62]   Long Chen, Martin Strauch, and Dorit Merhof. 2019. Instance Segmentation of Biomedical Images with an Object-aware Embedding Learned with Local Constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. https://doi.org/10.48550/ARXIV.2004.09821

[63]   Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[64]   Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[65]   Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv 2002.05709* (2020).

[66]   Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852* (2021).

[67]   Yuanyuan Chen, Boyang Li, Han Yu, Pengcheng Wu, and Chunyan Miao. 2021. Hydra: Hypergradient data relevance analysis for interpreting deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7081–7089.

[68]   Yi-Ting Chen, Xiaokai Liu, and Ming-Hsuan Yang. 2015. Multi-instance object segmentation with occlusion handling. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3470–3478. https://doi.org/10.1109/CVPR.2015.7298969

[69] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. 2020. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12472–12482. https://doi.org/10.1109/CVPR42600.2020.01249

[70] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. 2022. Instance-Dependent Label-Noise Learning with Manifold-Regularized Transition Matrix Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16630–16639.

[71] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.

[72] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. 2021. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *arXiv 2104.13840* (2021).

[73] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. 2012. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/459a4ddcb586f24efd9395aa7662bc7c-Paper.pdf

[74] Sergio Cubero, Won Suk Lee, Nuria Aleixos, Francisco Albert, and Jose Blasco. 2016. Automated systems based on machine vision for inspecting citrus fruits from the field to postharvest—a review. *Food and Bioprocess Technology* 9 (2016), 1623–1639.

[75] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. RandAugment: Practical automated data augmentation with a reduced search space. *arXiv 1909.13719* (2019).

[76] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. 2016. Instance-Sensitive Fully Convolutional Networks. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 534–549.

[77] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 764–773.

[78] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. 2021. CoAtNet: Marrying Convolution and Attention for All Data Sizes. *arXiv preprint arXiv:2106.04803* (2021).

[79] Jim Davis, Tong Liang, James Enouen, and Roman Ilin. 2019. Hierarchical semantic labeling with adaptive confidence. In *International Symposium on Visual Computing*. Springer, 169–183.

[80] Bert De Brabandere, Davy Neven, and Luc Van Gool. 2017. Semantic Instance Segmentation with a Discriminative Loss Function. CVPR 2017 Workshop on Deep Learning for Robotic Vision. https://arxiv.org/abs/1708.02551

[81] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*. 4690–4699.

[82] Dickson Despommier. 2010. *The vertical farm: feeding the world in the 21st century.* Macmillan.

[83] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv 1810.04805* (2019).

[84] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems* 31 (2018).

[85] Philipe A Dias, Amy Tabb, and Henry Medeiros. 2018. Multispecies fruit flower detection using a refined semantic segmentation network. *IEEE robotics and automation letters* 3, 4 (2018), 3003–3010.

[86] Philipe A. Dias, Amy Tabb, and Henry Medeiros. 2018. Multispecies Fruit Flower Detection Using a Refined Semantic Segmentation Network. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3003–3010. https://doi.org/10.1109/LRA.2018.2849498

[87] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. 2021. Local Temperature Scaling for Probability Calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6889–6899.

[88] Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. 2018. A Neural Multi-sequence Alignment TeCHnique (NeuMATCH). In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

[89] Tiago Domingues, Tomás Brandão, and João C Ferreira. 2022. Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. *Agriculture* 12, 9 (2022), 1350.

[90] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[91] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. CenterNet: Keypoint Triplets for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[92] Tom Duckett, Simon Pearson, Simon Blackmore, Bruce Grieve, Wen-Hua Chen, Grzegorz Cielniak, Jason Cleaversmith, Jian Dai, Steve Davis, Charles Fox, et al. 2018. Agricultural robotics: the future of robotic agriculture. *arXiv preprint arXiv:1806.06762* (2018).

[93] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341, 3 (2009), 1.

[94] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1915–1929. https://doi.org/10.1109/TPAMI.2012.231

[95] Konstantinos P Ferentinos. 2018. Deep learning models for plant disease detection and diagnosis. *Computers and electronics in agriculture* 145 (2018), 311–318.

[96] Roemi Fernández, Carlota Salinas, Héctor Montes, and Javier Sarria. 2014. Multisensory system for fruit harvesting robots. Experimental testing in natural scenarios and with different kinds of crops. *Sensors* 14, 12 (2014), 23885–23904.

[97] Ruth Fong and Andrea Vedaldi. 2018. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8730–8738.

[98] Alina Förster, Jens Behley, Jan Behmann, and Ribana Roscher. 2019. Hyperspectral plant disease forecasting using generative adversarial networks. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 1793–1796.

[99] Logan Frank, Christopher Wiegman, Jim Davis, and Scott Shearer. 2021. Confidence-Driven Hierarchical Classification of Cultivated Plant Stresses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2503–2512.

[100] Evan DG Fraser and Malcolm Campbell. 2019. Agriculture 5.0: reconciling production with planetary health. *One Earth* 1, 3 (2019), 278–280.

[101] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. 2017. DSSD : Deconvolutional Single Shot Detector. *arXiv Preprint 1701.06659* (2017).

[102] LiMin Fu. 1991. Rule Learning by Searching on Adapted Nets.. In *AAAI*, Vol. 91. 590–595.

[103] Alvaro Fuentes, Sook Yoon, Sang Cheol Kim, and Dong Sun Park. 2017. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* 17, 9 (2017), 2022.

[104] Pedro Galeano, Esdras Joseph, and Rosa E Lillo. 2015. The Mahalanobis distance for functional data with applications to classification. *Technometrics* 57, 2 (2015), 281–291.

[105] Damien Garreau and Dina Mardaoui. 2021. What does LIME really see in images?. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 3620–3629. https://proceedings.mlr.press/v139/garreau21a.html

[106] Damien Garreau and Ulrike von Luxburg. 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 1287–1296. https://proceedings.mlr.press/v108/garreau20a.html

[107] Yuanyue Ge, Ya Xiong, Gabriel Lins Tenorio, and Pål Johan From. 2019. Fruit Localization and Environment Perception for Strawberry Harvesting Robots. *IEEE Access* 7 (2019), 147642–147652. https://doi.org/10.1109/ACCESS.2019.2946369

[108] Jordi Gené-Mola, Ricardo Sanz-Cortiella, Joan R Rosell-Polo, Josep-Ramon Morros, Javier Ruiz-Hidalgo, Verónica Vilaplana, and Eduard Gregorio. 2020. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and Electronics in Agriculture* 169 (2020), 105165.

[109] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. 2021. Part-aware Panoptic Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5481–5490. https://doi.org/10.1109/CVPR46437.2021.00544

[110] Ross Girshick. 2015. Fast R-CNN. In *ICCV*.

[111] Juliana Freitas Santos Gomes and Fabiana Rodrigues Leta. 2012. Applications of computer vision techniques in the agriculture and food industry: a review. *European Food Research and Technology* 235 (2012), 989–1000.

[112] Andrea Gomez-Zavaglia, Juan Carlos Mejuto, and Jesus Simal-Gandara. 2020. Mitigation of emerging implications of climate change on food production systems. *Food Research International* 134 (2020), 109256.

[113] Stephen Gould, Richard Fulton, and Daphne Koller. 2009. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th International Conference on Computer Vision*. 1–8. https://doi.org/10.1109/ICCV.2009.5459211

[114] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2018. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv 1706.02677* (2018).

[115] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.

[116] Riccardo Gozzovelli, Benjamin Franchetti, Malik Bekmurat, and Fiora Pirri. 2021. Tip-burn stress detection of lettuce canopy grown in Plant Factories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1259–1268.

[117] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1321–1330.

[118] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. 2021. Latent-Optimized Adversarial Neural Transfer for Sarcasm Detection. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021)*.

[119] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. 2021. Latent-Optimized Adversarial Neural Transfer for Sarcasm Detection. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021)*. http://www.boyangli.org/paper/XuGuo-NAACL-2021.pdf

[120] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2022. Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16399–16409.

[121] Felix G Gustafson and Elnore Stoldt. 1936. Some relations between leaf area and fruit size in tomatoes. *Plant Physiology* 11, 2 (1936), 445.

[122] Salvador Gutiérrez, Alexander Wendel, and James Underwood. 2019. Ground based hyperspectral imaging for extensive mango yield estimation. *Computers and Electronics in Agriculture* 157 (2019), 126–135.

[123] Md Tarek Habib, Md Ariful Islam Arif, Sumaita Binte Shorif, Mohammad Shorif Uddin, and Farruk Ahmed. 2021. Machine Vision-Based Fruit and Vegetable Disease Recognition: A Review. *Computer Vision and Machine Learning in Agriculture* (2021), 143–157.

[124] Michael Halstead, Christopher McCool, Simon Denman, Tristan Perez, and Clinton Fookes. 2018. Fruit quantity and ripeness estimation using a robotic vision system. *IEEE robotics and automation LETTERS* 3, 4 (2018), 2995–3002.

[125] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 5679–5690.

[126] Nicolai Häni, Pravakar Roy, and Volkan Isler. 2020. MinneApple: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters* 5, 2 (2020), 852–858.

[127] X Hao, X Guo, J Zheng, L Celeste, S Kholsa, and X Chen. 2015. Response of greenhouse tomato to different vertical spectra of LED lighting under overhead high pressure sodium and plasma lighting. In *International Symposium on New Technologies and Management for Greenhouses-GreenSys2015 1170*. 1003–1110.

[128] Xiuming Hao and Athanasios P Papadopoulos. 1999. Effects of supplemental lighting and cover materials on growth, photosynthesis, biomass partitioning, early yield and quality of greenhouse cucumber. *Scientia Horticulturae* 80, 1-2 (1999), 1–18.

[129] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2014. Simultaneous Detection and Segmentation. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 297–312.

[130] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M. Dai, and Dustin Tran. 2021. Training independent subnetworks for robust prediction. In *ICLR*.

[131] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. 2016. Boundary-aware Instance Segmentation. *arXiv Preprint 1612.03129* (2016). https://doi.org/10.48550/ARXIV.1612.03129

[132] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *ICCV*.

[133] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2018. Mask R-CNN. *arXiv 1703.06870* (2018).

[134] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv 1512.03385* (2015).

[135] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015). arXiv:1512.03385 [cs.CV]

[136] Leilei He, Wentai Fang, Guanao Zhao, Zhenchao Wu, Longsheng Fu, Rui Li, Yaqoob Majeed, and Jaspreet Dhupia. 2022. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. *Computers and Electronics in Agriculture* 195 (2022), 106812.

[137] Katherine L. Hermann and Andrew K. Lampinen. 2020. What shapes feature representations? Exploring datasets, architectures, and training. *arXiv2006.12433* (2020).

[138] Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb

[139] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv 1704.04861* (2017).

[140] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv PrePrint 1704.04861* (2017).

[141] Chunhua Hu, Xuan Liu, Zhou Pan, and Pingping Li. 2019. Automatic detection of single ripe tomato on plant combining faster R-CNN and intuitionistic fuzzy set. *IEEE Access* 7 (2019), 154683–154696.

[142] Xiaoping Huang, Zelin Hu, Xiaorun Wang, Xuanjiang Yang, Jian Zhang, and Daoling Shi. 2019. An Improved Single Shot Multibox Detector Method Applied in Body Condition Score for Dairy Cows. *Animals* 9, 7 (2019). https://doi.org/10.3390/ani9070470

[143] Yo-Ping Huang, Tzu-Hao Wang, and Haobijam Basanta. 2020. Using fuzzy mask R-CNN model to automatically identify tomato ripeness. *IEEE Access* 8 (2020), 207672–207682.

[144] Wallace E Huffman. 2012. The status of labor-saving mechanization in US fruit and vegetable harvesting. *Choices* 27, 316-2016-6262 (2012).

[145] David Hughes, Marcel Salathé, et al. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* (2015).

[146] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* 110, 3 (2021), 457–506.

[147] Sarah Ibrahimi, Arnaud Sors, Rafael Sampaio de Rezende, and Stéphane Clinchant. 2022. Learning with label noise for image retrieval by selecting interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2181–2190.

[148] Joana Iljazi. 2017. Deep learning for image-based prediction of plant growth in City Farms. (2017).

[149] Zahid Iqbal, Muhammad Attique Khan, Muhammad Sharif, Jamal Hussain Shah, Muhammad Habib ur Rehman, and Kashif Javed. 2018. An automated detection and classification of citrus plant diseases using image processing techniques: A review. *Computers and electronics in agriculture* 153 (2018), 12–32.

[150] David Ireri, Eisa Belal, Cedric Okinda, Nelson Makange, and Changying Ji. 2019. A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. *Artificial Intelligence in Agriculture* 2 (2019), 28–37.

[151] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. 2018. i-RevNet: Deep Invertible Networks. In *International Conference on Learning Representations (ICLR)*.

[152] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795* (2021).

[153] Lu Jiang, Mason Liu Di Huang, and Weilong Yang. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*.

[154] Vijay Kakani, Van Huan Nguyen, Basivi Praveen Kumar, Hakil Kim, and Visweswara Rao Pasupuleti. 2020. A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research* 2 (2020), 100033.

[155] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly Supervised Concept Expansion for General Purpose Vision Models. *arXiv preprint arXiv:2202.02317* (2022).

[156] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. 2020. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization.. In *IJCAI*. 2855–2862.

[157] Ramesh Kestur, Avadesh Meduri, and Omkar Narasipura. 2019. MangoNet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Engineering Applications of Artificial Intelligence* 77 (2019), 59–69.

[158] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. 2019. Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[159] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[160] Anand Koirala, KB Walsh, Zhenglin Wang, and C McCarthy. 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture* 20, 6 (2019), 1107–1135.

[161] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. 2020. FoveaBox: Beyond Anchor-Based Object Detection. *IEEE Transactions on Image Processing* 29 (2020), 7389–7398. https://doi.org/10.1109/TIP.2020.3002345

[162] Dean A Kopsell, Carl E Sams, and Robert C Morrow. 2015. Blue wavelengths from LED lighting increase nutritionally important metabolites in specialty crops. *HortScience* 50, 9 (2015), 1285–1288.

[163] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. 2020. SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems* 203 (2020), 106164. https://doi.org/10.1016/j.knosys.2020.106164

[164] R Krishnamurthy. 2014. vertical farming: Singapore's Solution to feed the local urban Population. *Permaculture Research Institute* (2014).

[165] Alex Krizhevsky. 2014. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997* (2014).

[166] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* (Lake Tahoe, Nevada) *(NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[167] Anders Krogh Mortensen, Soren Skovsen, Henrik Karstoft, and Rene Gislum. 2019. The Oil Radish Growth Dataset for Semantic Segmentation and Yield Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[168] Ferhat Kurtulmus, Won Suk Lee, and Ali Vardar. 2014. Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network. *Precision agriculture* 15, 1 (2014), 57–79.

[169] Keerthy Kusumam, Tomáš Krajník, Simon Pearson, Tom Duckett, and Grzegorz Cielniak. 2017. 3D-vision based detection, localization, and sizing of broccoli heads in the field. *Journal of Field Robotics* 34, 8 (2017), 1505–1518.

[170] L'ubor Ladický, Chris Russell, Pushmeet Kohli, and Philip H.S. Torr. 2009. Associative hierarchical CRFs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*. 739–746. https://doi.org/10.1109/ICCV.2009.5459248

[171] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.

[172] Peter D Lancashire, Hermann Bleiholder, T van den Boom, P Langelüddeke, Reinhold Stauss, Elfriede Weber, and A Witzenberger. 1991. A uniform decimal code for growth stages of crops and weeds. *Annals of applied Biology* 119, 3 (1991), 561–601.

[173] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2017. FractalNet: Ultra-Deep Neural Networks without Residuals. In *ICLR*.

[174] Hei Law and Jia Deng. 2018. CornerNet: Detecting Objects as Paired Keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[175] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521 (2015), 436–444.

[176] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[177] Joon-Woo Lee, Taewon Moon, and Jung-Eek Son. 2021. Development of Growth Estimation Algorithms for Hydroponic Bell Peppers Using Recurrent Neural Networks. *Horticulturae* 7, 9 (2021). https://doi.org/10.3390/horticulturae7090284

[178] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[179] Han Li, Won Suk Lee, and Ku Wang. 2016. Immature green citrus fruit detection and counting based on fast normalized cross correlation (FNCC) using natural outdoor colour images. *Precision Agriculture* 17, 6 (2016), 678–697.

[180] Kai Li, Martin Renqiang Min, and Yun Fu. 2019. Rethinking zero-shot learning: A conditional visual classification perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3583–3592.

[181] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. 2019. Attention-Guided Unified Network for Panoptic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7019–7028. https://doi.org/10.1109/CVPR.2019.00719

[182] Yuanzhi Li and Yingyu Liang. 2018. Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf

[183] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. [n. d.]. Fully Convolutional Instance-Aware Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[184] Yanfen Li, Hanxiang Wang, L Minh Dang, Abolghasem Sadeghi-Niaraki, and Hyeonjoon Moon. 2020. Crop pest recognition in natural scenes using convolutional neural networks. *Computers and Electronics in Agriculture* 169 (2020), 105174.

[185] Yang Li and Jiachen Yang. 2021. Meta-learning baselines and database for few-shot classification in agriculture. *Computers and Electronics in Agriculture* 182 (2021), 106055.

[186] Zhizhong Li and Derek Hoiem. 2020. Improving Confidence Estimates for Unfamiliar Examples. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2683–2692. https://doi.org/10.1109/CVPR42600.2020.00276

[187] Guichao Lin, Yunchao Tang, Xiangjun Zou, Juntao Xiong, and Yamei Fang. 2020. Color-, depth-, and shape-based 3D fruit detection. *Precision Agriculture* 21, 1 (2020), 1–17.

[188] Guichao Lin, Yunchao Tang, Xiangjun Zou, Juntao Xiong, and Jinhui Li. 2019. Guava detection and pose estimation using a low-cost RGB-D sensor in the field. *Sensors* 19, 2 (2019), 428.

[189] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[190] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[191] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. 2021. Noise-resistant Deep Metric Learning with Ranking-based Instance Selection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. http://www.boyangli.org/paper/ChangLiu-CVPR-2021.pdf

[192] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. 2019. An End-To-End Network for Panoptic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6165–6174. https://doi.org/10.1109/CVPR.2019.00633

[193] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective. In *CVPR*. 2970–2979.

[194] Qiuhua Liu, Xuejun Liao, and Lawrence Carin. 2007. Semi-Supervised Multitask Learning. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/a34bacf839b923770b2c360eefa26748-Paper.pdf

[195] Tian-Hu Liu, Reza Ehsani, Arash Toudeshki, Xiang-Jun Zou, and Hong-Jun Wang. 2018. Detection of citrus fruit and tree trunks in natural environments using a multi-elliptical boundary model. *Computers in Industry* 99 (2018), 9–16.

[196] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 21–37.

[197] Yang Liu, Duolin Wang, Fei He, Juexin Wang, Trupti Joshi, and Dong Xu. 2019. Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Frontiers in Genetics* 10 (2019), 1091. https://doi.org/10.3389/fgene.2019.01091

[198] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10012–10022.

[199] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR).* 3431–3440.

[200] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. *arXiv preprint arXiv:2206.08916* (2022).

[201] Jie-Yan Lu, Chung-Liang Chang, and Yan-Fu Kuo. 2019. Monitoring growth rate of lettuce using deep convolutional neural networks. In *2019 ASABE Annual International Meeting.* American Society of Agricultural and Biological Engineers, 1.

[202] Shilei Lyu, Yawen Zhao, Ruiyao Li, Zhen Li, Renjie Fan, and Qiafeng Li. 2022. Embedded Sensing System for Recognizing Citrus Flowers Using Cascaded Fusion YOLOv4-CF+ FPGA. *Sensors* 22, 3 (2022), 1255.

[203] Juncheng Ma, Keming Du, Lingxian Zhang, Feixiang Zheng, Jinxiang Chu, and Zhongfu Sun. 2017. A segmentation method for greenhouse vegetable foliar disease spots images using color information and region growing. *Computers and Electronics in Agriculture* 142 (2017), 110–117.

[204] Juncheng Ma, Keming Du, Feixiang Zheng, Lingxian Zhang, Zhihong Gong, and Zhongfu Sun. 2018. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. *Computers and electronics in agriculture* 154 (2018), 18–24.

[205] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 5188–5196.

[206] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. 2020. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging* 39, 12 (2020), 3868–3878. https://doi.org/10.1109/TMI.2020.3006437

[207] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A Tsaftaris. 2016. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters* 81 (2016), 80–89.

[208] Lj Miranda. [n. d.]. *Towards data-centric machine learning: a short review.* https://ljvmiranda921.github.io/notebook/2021/07/30/data-centric-ml/

[209] Christoph Molnar. 2020. *Interpretable machine learning.* https://christophm.github.io/interpretable-ml-book/

[210] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), 193–209.

[211] Jednipat Moonrinta, Supawadee Chaivivatrakul, Matthew N Dailey, and Mongkol Ekpanyapong. 2010. Fruit detection, tracking, and 3D reconstruction for crop mapping and yield estimation. In *2010 11th International Conference on Control Automation Robotics & Vision.* IEEE, 1181–1186.

[212] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. 2015. Inceptionism: Going deeper into neural networks. (2015).

[213] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating Deep Neural Networks using Focal Loss. In *Advances in Neural Information Processing Systems,* H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 15288–15299. https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf

[214] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?. In *Advances in Neural Information Processing Systems,* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

[215] Joanna M Nassar, Sherjeel M Khan, Diego Rosas Villalva, Maha N Nour, Amani S Almuslem, and Muhammad M Hussain. 2018. Compliant plant wearables for localized microclimate and plant growth monitoring. *npj Flexible Electronics* 2, 1 (2018), 1–12.

[216] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. 2019. Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 8829–8837. https://doi.org/10.1109/CVPR.2019.00904

[217] Andrew Ng. [n. d.]. *A Chat with Andrew on MLOps: From Model-centric to Data-centric AI.* https://www.youtube.com/watch?v=06-AZXmwHjo

[218] Canh Nguyen, Vasit Sagan, Matthew Maimaitiyiming, Maitiniyazi Maimaitijiang, Sourav Bhadra, and Misha T. Kwasniewski. 2021. Early Detection of Plant Viral Disease Using Hyperspectral Imaging and Deep Learning. *Sensors* 21, 3 (2021). https://doi.org/10.3390/s21030742

[219] Xueping Ni, Changying Li, Huanyu Jiang, and Fumiomi Takeda. 2020. Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield. *Horticulture research* 7 (2020).

[220] Sai Vidyaranya Nuthalapati and Anirudh Tunga. 2021. Multi-domain few-shot learning and dataset for agricultural applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1399–1408.

[221] Emmanuel Karlo Nyarko, Ivan Vidović, Kristijan Radočaj, and Robert Cupec. 2018. A nearest neighbor approach for fruit recognition in RGB-D images based on detection of convex surfaces. *Expert Systems with Applications* 114 (2018), 454–466.

[222] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 685–694.

[223] Ahmad Ostovar, Ola Ringdahl, and Thomas Hellström. 2018. Adaptive image thresholding of yellow peppers for a harvesting robot. *Robotics* 7, 1 (2018), 11.

[224] Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9, 1 (1979), 62–66.

[225] Christian Payer, Darko Štern, Thomas Neff, Horst Bischof, and Martin Urschler. 2018. Instance Segmentation and Tracking with Cosine Embeddings and Recurrent Hourglass Networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*

[226] Tejaswini Pedapati, Avinash Balakrishnan, Karthikeyan Shanmugam, and Amit Dhurandhar. 2020. Learning global transparent models consistent with local contrastive explanations. *Advances in neural information processing systems* 33 (2020), 3592–3602.

[227] Yao Peng, Mary M Dallas, José T Ascencio-Ibáñez, J Steen Hoyer, James Legg, Linda Hanley-Bowdoin, Bruce Grieve, and Hujun Yin. 2022. Early detection of plant virus infection using multispectral imaging and spatial–spectral machine learning. *Scientific Reports* 12, 1 (2022), 3113.

[228] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. *arXiv 1701.06548* (2017).

[229] Wilhelm Pfeffer. 1900. *The physiology of plants: a treatise upon the metabolism and sources of energy in plants.* Vol. 1. Clarendon Press.
[230] Pedro O Pinheiro and Ronan Collobert. 2015. Doll´ ar P. Learning to segment object candidates. In *Proc. the 28th Int. Conf. Neural Information Processing Systems.* 1990–1998.
[231] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. 2016. Learning to Refine Object Segments. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 75–91.
[232] Przemyslaw Prusinkiewicz. 2002. Art and science of life: designing and growing virtual plants with L-systems. In *XXVI International Horticultural Congress: Nursery Crops; Development, Evaluation, Production and Use 630.* 15–28.
[233] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems* 33 (2020), 19920–19930.
[234] Zhongang Qi, Saeed Khorram, and Li Fuxin. 2021. Embedding deep networks into visual explanations. *Artificial Intelligence* 292 (2021), 103435.
[235] Redmond R Shamshiri, Cornelia Weltzien, Ibrahim A Hameed, Ian J Yule, Tony E Grift, Siva K Balasundram, Lenka Pitonakova, Desa Ahmad, and Girish Chowdhary. 2018. Research and development in agricultural robotics: A perspective of digital farming. (2018).
[236] K Ragazou, A Garefalakis, E Zafeiriou, and I Passas. 2022. Agriculture 5.0: A New Strategic Management Mode for a Cut Cost and an Energy Efficient Agriculture Sector. Energies 2022, 15, 3113.
[237] Parastoo Rahimi, Md Saiful Islam, Phelipe Magalhães Duarte, Sina Salajegheh Tazerji, Md Abdus Sobur, Mohamed E El Zowalaty, Hossam M Ashour, and Md Tanvir Rahman. 2021. Impact of the COVID-19 pandemic on food production and animal health. *Trends in Food Science & Technology* (2021).
[238] Maryam Rahnemoonfar and Clay Sheppard. 2017. Deep count: fruit counting based on deep simulated learning. *Sensors* 17, 4 (2017), 905.
[239] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. *arXiv 1506.02640* (2016).
[240] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
[241] Tanzeel U Rehman, Md Sultan Mahmud, Young K Chang, Jian Jin, and Jaemyung Shin. 2019. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and electronics in agriculture* 156 (2019), 585–605.
[242] Mengye Ren and Richard S Zemel. 2017. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 6656–6664.
[243] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems.*
[244] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv 1506.01497* (2016).
[245] A. Reyes-Yanes, P. Martinez, and R. Ahmad. 2020. Real-time growth rate and fresh weight estimation for little gem romaine lettuce in aquaponic grow beds. *Computers and Electronics in Agriculture* 179 (2020), 105827.   https://doi.org/10.1016/j.compag.2020.105827
[246] A Reyes-Yanes, Pablo Martinez, and R Ahmad. 2020. Real-time growth rate and fresh weight estimation for little gem romaine lettuce in aquaponic grow beds. *Computers and Electronics in Agriculture* 179 (2020), 105827.
[247] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 1135–1144.
[248] Joe M Roberts, Toby JA Bruce, James M Monaghan, Tom W Pope, Simon R Leather, and Andrew M Beacham. 2020. Vertical farming systems bring new considerations for pest and disease management. *Annals of Applied Biology* 176, 3 (2020), 226–232.
[249] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling Climate Change with Machine Learning. *ACM Comput. Surv.* 55, 2, Article 42 (feb 2022), 96 pages.   https://doi.org/10.1145/3485128
[250] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. 2017. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems* 19, 1 (2017), 263–272.
[251] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. 2016. Recurrent instance segmentation. In *European conference on computer vision.* Springer, 312–329.
[252] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.
[253] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.   https://doi.org/10.1007/s11263-015-0816-y
[254] Radu Bogdan Rusu. 2010. Semantic 3D object maps for everyday manipulation in human living environments. *KI-Künstliche Intelligenz* 24, 4 (2010), 345–348.
[255] Inkyu Sa, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. 2016. Deepfruits: A fruit detection system using deep neural networks. *sensors* 16, 8 (2016), 1222.
[256] Inkyu Sa, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 8 (2016).   https://doi.org/10.3390/s16081222
[257] Verónica Saiz-Rubio and Francisco Rovira-Más. 2020. From smart farming towards agriculture 5.0: A review on crop data management. *Agronomy* 10, 2 (2020), 207.
[258] Amaia Salvador, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i Nieto. 2017. Recurrent Neural Networks for Semantic Instance Segmentation. *arXiv Preprint 1712.00617* (2017).   https://arxiv.org/abs/1712.00617
[259] Amaia Salvador, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i Nieto. 2017. Recurrent neural networks for semantic instance segmentation. *arXiv preprint arXiv:1712.00617* (2017).

[260] Hanno Scharr, Massimo Minervini, Andrew P French, Christian Klukas, David M Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, et al. 2016. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications* 27, 4 (2016), 585–606.

[261] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. 2021. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access* 9 (2021), 82146–82168.

[262] Michael Gomez Selvaraj, Alejandro Vergara, Henry Ruiz, Nancy Safari, Sivalingam Elayabalan, Walter Ocimati, and Guy Blomme. 2019. AI-powered banana diseases and pest detection. *Plant Methods* 15, 1 (2019), 1–11.

[263] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[264] Woo Chaw Seng and Seyed Hadi Mirisaee. 2009. A new method for fruits recognition system. In *2009 international conference on electrical engineering and informatics*, Vol. 1. IEEE, 130–134.

[265] Jayavelu Senthilnath, Akanksha Dokania, Manasa Kandukuri, KN Ramesh, Gautham Anand, and SN Omkar. 2016. Detection of tomatoes using spectral-spatial methods in remotely sensed RGB images captured by UAV. *Biosystems engineering* 146 (2016), 16–32.

[266] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).

[267] Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. Learning from the Best: Rationalizing Predictions by Adversarial Information Calibration.. In *AAAI*. 13771–13779.

[268] Manish Sharma, Mayur Dhanaraj, Srivallabha Karnam, Dimitris G Chachlakis, Raymond Ptucha, Panos P Markopoulos, and Eli Saber. 2020. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), 1497–1508.

[269] Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*. Springer, 467–482.

[270] Rui Shi, Tianxing Li, and Yasushi Yamaguchi. 2020. An attribution-based pruning method for real-time mango detection with YOLO network. *Computers and electronics in agriculture* 169 (2020), 105214.

[271] Ruifeng Shi, Deming Zhai, Xianming Liu, Junjun Jiang, and Wen Gao. 2020. Rectified meta-learning from noisy labels for robust image-based plant disease diagnosis. *arXiv preprint arXiv:2003.07603* (2020).

[272] Shigeharu Shimamura. [n. d.]. Indoor Cultivation for the Future. https://frc.ri.cmu.edu/~ssingh/VF/Challenges_in_Vertical_Farming/Schedule_files/SHIMAMURA.pdf.

[273] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems* 34 (2021), 11352–11363.

[274] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Interpretable deep learning by propagating activation differences. *arXiv preprint arXiv:1605.01713* 4 (2016).

[275] Singh Chouhan Siddharth, Kaul Ajay, Pratap Singh Uday, and Jain Sanjeev. 2019. A database of leaf images: practice towards plant conservation with plant pathology. *Mendeley Data* (2019).

[276] Aleksandra Sidor and Piotr Rzymski. 2020. Dietary choices and habits during COVID-19 lockdown: experience from Poland. *Nutrients* 12, 6 (2020), 1657.

[277] Claudênia Ferreira da Silva, Carlos Hidemi Uesugi, Luiz Eduardo Bassay Blum, Abi Soares dos Anjos Marques, and Marisa Álvares da Silva Velloso Ferreira. 2016. Molecular detection of Erwinia psidii in guava plants under greenhouse and field conditions. *Ciência Rural* 46 (2016), 1528–1534.

[278] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550 (2017), 354–359.

[279] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[280] Uday Pratap Singh, Siddharth Singh Chouhan, Sukirty Jain, and Sanjeev Jain. 2019. Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. *IEEE Access* 7 (2019), 43721–43729.

[281] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[282] Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. 2021. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176* (2021).

[283] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[284] Jie Song, Chengchao Shen, Jie Lei, An-Xiang Zeng, Kairi Ou, Dacheng Tao, and Mingli Song. 2018. Selective zero-shot classification with augmented attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 468–483.

[285] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. 2018. Transductive unbiased embedding for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1024–1033.

[286] K Sornalakshmi, G Sujatha, S Sindhu, and D Hemavathi. 2022. A Technical Survey on Deep Learning and AI Solutions for Plant Quality and Health Indicators Monitoring in Agriculture. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 984–988.

[287] Edgar P Spalding and Nathan D Miller. 2013. Image analysis is driving a renaissance in growth measurement. *Current opinion in plant biology* 16, 1 (2013), 100–104.

[288] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning?. In *International Conference on Machine Learning*. PMLR, 9120–9132.

[289] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2022. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=4nPswr1KcP

[290] Richard N Strange and Peter R Scott. 2005. Plant disease: a threat to global food security. *Annual review of phytopathology* 43, 1 (2005), 83–116.

[291] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[292] Jun Sun, Xiaofei He, Xiao Ge, Xiaohong Wu, Jifeng Shen, and Yingying Song. 2018. Detection of key organs in tomato based on deep migration learning in a complex background. *Agriculture* 8, 12 (2018), 196.

[293] Kaiqiong Sun, Xuan Wang, Shoushuai Liu, and ChangHua Liu. 2021. Apple, peach, and pear flower detection using semantic segmentation network and shape constraint level set. *Computers and Electronics in Agriculture* 185 (2021), 106150.

[294] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 403–412.

[295] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[296] Nik Susič, Uroš Žibrat, Saša Širca, Polona Strajnar, Jaka Razinger, Matej Knapič, Andrej Vončina, Gregor Urek, and Barbara Gerič Stare. 2018. Discrimination between abiotic and biotic drought stress in tomatoes using hyperspectral imaging. *Sensors and actuators B: Chemical* 273 (2018), 842–852.

[297] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv Preprint 1512.00567* (2015).

[298] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[299] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[300] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, and Peter Corke. 2018. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research* 37, 4-5 (2018), 405–420.

[301] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

[302] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*.

[303] Wenzhi Tang, Tingting Yan, Fei Wang, Jingxian Yang, Jian Wu, Jianlong Wang, Tianli Yue, and Zhonghong Li. 2019. Rapid fabrication of wearable carbon nanotube/graphite strain sensor for real-time monitoring of plant growth. *Carbon* 147 (2019), 295–302.

[304] Nima Teimouri, Mads Dyrmann, Per Rydahl Nielsen, Solvejg Kopp Mathiassen, Gayle J. Somerville, and Rasmus Nyholm Jørgensen. 2018. Weed Growth Stage Estimator Using Deep Convolutional Neural Networks. *Sensors* 18, 5 (2018). http://www.mdpi.com/1424-8220/18/5/1580

[305] Mercè Teixidó, Davinia Font, Tomàs Pallejà, Marcel Tresanchez, Miquel Nogués, and Jordi Palacín. 2012. Definition of linear color models in the RGB vector color space to detect red peaches in orchard images taken under natural illumination. *Sensors* 12, 6 (2012), 7701–7718.

[306] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/36ad8b5f42db492827016448975cc22d-Paper.pdf

[307] Hongkun Tian, Tianhai Wang, Yadong Liu, Xi Qiao, and Yanzhou Li. 2020. Computer vision technology in agricultural automation—A review. *Information Processing in Agriculture* 7, 1 (2020), 1–19.

[308] Mengxiao Tian, Hao Guo, Hong Chen, Qing Wang, Chengjiang Long, and Yuhao Ma. 2019. Automated pig counting using deep learning. *Computers and Electronics in Agriculture* 163 (2019), 104840. https://doi.org/10.1016/j.compag.2019.05.049

[309] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[310] Julio Torres-Tello and Seok-Bum Ko. 2022. Optimizing a Multispectral-Images-Based DL model, through feature selection, pruning and quantization. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1352–1356.

[311] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. 2021. Going Deeper With Image Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 32–42.

[312] Mukesh Kumar Tripathi and Dhananjay D Maktedar. 2020. A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A survey. *Information Processing in Agriculture* 7, 2 (2020), 183–203.

[313] Magdalena Trojak, Ernest Skowron, Tomasz Sobala, Maciej Kocurek, and Jan Pałyga. 2022. Effects of partial replacement of red by green light in the growth spectrum on photomorphogenesis and photosynthesis in tomato plants. *Photosynthesis research* 151, 3 (2022), 295–312.

[314] Jordan Ubbens, Mikolaj Cieslak, Przemyslaw Prusinkiewicz, and Ian Stavness. 2018. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant methods* 14, 1 (2018), 1–10.

[315] Nathalie van Wijkvliet. [n. d.]. No space, no problem. How Singapore is turning into an edible paradise. https://sustainableurbandelta.com/singapore-30-by-30-food-system/. Accessed: 2022-8-15.

[316] Charles Veys, Fokion Chatziavgerinos, Ali AlSuwaidi, James Hibbert, Mark Hansen, Gytis Bernotas, Melvyn Smith, Hujun Yin, Stephen Rolfe, and Bruce Grieve. 2019. Multispectral imaging for presymptomatic analysis of light leaf spot in oilseed rape. *Plant methods* 15 (2019), 1–12.

[317] Adar Vit and Guy Shani. 2018. Comparing rgb-d sensors for close range outdoor agricultural phenotyping. *Sensors* 18, 12 (2018), 4413.

[318] Fabio Vulpi, Roberto Marani, Antonio Petitti, Giulio Reina, and Annalisa Milella. 2022. An RGB-D multi-view perspective for autonomous agricultural robots. *Computers and Electronics in Agriculture* 202 (2022), 107419.

[319] Dong Wang and Xiaoyang Tan. 2017. Robust distance metric learning via Bayesian inference. *IEEE Transactions on Image Processing* 27, 3 (2017), 1542–1553.

[320] Dongyi Wang, Robert Vinson, Maxwell Holmes, Gary Seibel, Avital Bechar, Shimon Nof, and Yang Tao. 2019. Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN). *Scientific reports* 9, 1 (2019),

1717   1–14.

[321] Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. 2021. Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 11809–11820. https://proceedings.neurips.cc/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf

[322] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*. 5265–5274.

[323] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052* (2022).

[324] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

[325] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. 2020. SOLO: Segmenting Objects by Locations. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 649–665.

[326] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. 2020. SOLOv2: Dynamic and Fast Instance Segmentation. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17721–17732. https://proceedings.neurips.cc/paper/2020/file/cd3afef9b8b89558cd56638c3631868a-Paper.pdf

[327] Yulong Wang, Hang Su, Bo Zhang, and Xiaolin Hu. 2018. Interpret neural networks by identifying critical data routing paths. In *proceedings of the IEEE conference on computer vision and pattern recognition*. 8906–8914.

[328] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[329] Zhenglin Wang, Kerry Walsh, and Anand Koirala. 2019. Mango fruit load estimation using a video based MangoYOLO—Kalman filter—hungarian algorithm method. *Sensors* 19, 12 (2019), 2742.

[330] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*. 13726–13735.

[331] Xiangqin Wei, Kun Jia, Jinhui Lan, Yuwei Li, Yiliang Zeng, and Chunmei Wang. 2014. Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot. *Optik* 125, 19 (2014), 5684–5689.

[332] Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning. In *ICLR*.

[333] Jan Weyler, Federico Magistri, Peter Seitz, Jens Behley, and Cyrill Stachniss. 2022. In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2725–2734.

[334] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2021. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. (2021). https://doi.org/10.48550/ARXIV.2112.06409

[335] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. 2022. Sparse Object-Level Supervision for Instance Segmentation With Pixel Embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4402–4411.

[336] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. 2022. Sparse object-level supervision for instance segmentation with pixel embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4402–4411.

[337] Arissa Wongpanich, Hieu Pham, James Demmel, Mingxing Tan, Quoc Le, Yang You, and Sameer Kumar. 2021. Training EfficientNets at Super-computer Scale: 83% ImageNet Top-1 Accuracy in One Hour. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 947–950.

[338] Jingui Wu, Baohua Zhang, Jun Zhou, Yingjun Xiong, Baoxing Gu, and Xiaolong Yang. 2019. Automatic recognition of ripening tomatoes by combining multi-feature fusion with a bi-layer classification strategy for harvesting robots. *Sensors* 19, 3 (2019), 612.

[339] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. 2019. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8787–8796.

[340] Yuli Wu, Long Chen, and Dorit Merhof. 2020. Improving Pixel Embedding Learning through Intermediate Distance Regression Supervision for Instance Segmentation. In *European Conference on Computer Vision Workshop*. Springer, 213–227.

[341] Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. 2019. Logan: Latent optimisation for generative adversarial networks. *arXiv preprint arXiv:1912.00953* (2019).

[342] Adnelba Vitória Oliveira Xavier, Geovani Soares de Lima, Hans Raj Gheyi, André Alisson Rodrigues da Silva, Lauriane Almeida dos Anjos Soares, and Cassiano Nogueira de Lacerda. 2022. Gas exchange, growth and quality of guava seedlings under salt stress and salicylic acid. *Revista Ambiente & Água* 17 (2022).

[343] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 69–77.

[344] Jingjing Xie, Bing Xu, and Zhang Chuang. 2013. Horizontal and Vertical Ensemble with Deep Representation for Classification. *arXiv 1306.2759* (2013).

[345] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv Preprint 1611.05431* (2016).

[346] Haotian Yan, Zhe Li, Weijian Li, Changhu Wang, Ming Wu, and Chuang Zhang. 2021. ConTNet: Why not use convolution and transformer at the same time? *arXiv Preprint 2104.13497* (2021).

[347] Biyun Yang and Yong Xu. 2021. Applications of deep-learning approaches in horticultural research: a review. *Horticulture Research* 8 (2021).

[348] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3081–3089.

[349] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems* 32 (2019).

[350] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems* 31 (2018).

[351] T Yeshitela, PJ Robbertse, and PJC Stassen. 2005. Effects of pruning on flowering, yield and fruit quality in mango (Mangifera indica). *Australian Journal of Experimental Agriculture* 45, 10 (2005), 1325–1330.

[352] Michael Yeung, Leonardo Rundo, Yang Nan, Evis Sala, Carola-Bibiane Schönlieb, and Guang Yang. 2021. Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. *arXiv preprint arXiv:2111.00528* (2021).

[353] Hui Ying, Zhaojin Huang, Shu Liu, Tianjia Shao, and Kun Zhou. 2021. EmbedMask: Embedding Coupling for Instance Segmentation.. In *IJCAI*. 1266–1273.

[354] Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems* 34 (2021), 12822–12835.

[355] Yang Yu, Kailiang Zhang, Hui Liu, Li Yang, and Dongxing Zhang. 2020. Real-Time Visual Localization of the Picking Points for a Ridge-Planting Strawberry Harvesting Robot. *IEEE Access* 8 (2020), 116556–116568. https://doi.org/10.1109/ACCESS.2020.3003034

[356] Yang Yu, Kailiang Zhang, Li Yang, and Dongxing Zhang. 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture* 163 (2019), 104846.

[357] Zhiwen Yu, Hau-San Wong, and Guihua Wen. 2011. A modified support vector machine and its application to image segmentation. *Image and Vision Computing* 29, 1 (2011), 29–40.

[358] Hongbo Yuan, Jiajun Zhu, Qifan Wang, Man Cheng, and Zhenjiang Cai. 2022. An Improved DeepLab v3+ Deep Learning Network Applied to the Segmentation of Grape Leaf Black Rot Spots. *Frontiers in Plant Science* 13 (2022).

[359] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. 2021. Incorporating Convolution Designs Into Visual Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 579–588.

[360] Ting Yuan, Lin Lv, Fan Zhang, Jun Fu, Jin Gao, Junxiong Zhang, Wei Li, Chunlong Zhang, and Wenqiang Zhang. 2020. Robust cherry tomatoes detection algorithm in greenhouse scene based on SSD. *Agriculture* 10, 5 (2020), 160.

[361] Ilaria Zambon, Massimo Cecchini, Gianluca Egidi, Maria Grazia Saporito, and Andrea Colantoni. 2019. Revolution 4.0: Industry vs. agriculture in a future development for SMEs. *Processes* 7, 1 (2019), 36.

[362] Baohua Zhang, Yuanxin Xie, Jun Zhou, Kai Wang, and Zhen Zhang. 2020. State-of-the-art robotic grippers, grasping and control strategies, as well as their applications in agricultural robots: A review. *Computers and Electronics in Agriculture* 177 (2020), 105694.

[363] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. *arXiv 1611.03530* (2017).

[364] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[365] Li Zhang, Guan Gui, Abdul Mateen Khattak, Minjuan Wang, Wanlin Gao, and Jingdun Jia. 2019. Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot. *IEEE Access* 7 (2019), 56028–56038.

[366] Li Zhang, Jingdun Jia, Guan Gui, Xia Hao, Wanlin Gao, and Minjuan Wang. 2018. Deep Learning Based Improved Classification System for Designing Tomato Harvesting Robot. *IEEE Access* 6 (2018), 67940–67950. https://doi.org/10.1109/ACCESS.2018.2879324

[367] Li Zhang, Jingdun Jia, Guan Gui, Xia Hao, Wanlin Gao, and Minjuan Wang. 2018. Deep learning based improved classification system for designing tomato harvesting robot. *IEEE Access* 6 (2018), 67940–67950.

[368] Lingxian Zhang, Zanyu Xu, Dan Xu, Juncheng Ma, Yingyi Chen, and Zetian Fu. 2020. Growth monitoring of greenhouse lettuce based on a convolutional neural network. *Horticulture research* 7 (2020).

[369] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. 2020. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[370] Shanwen Zhang, Subing Zhang, Chuanlei Zhang, Xianfeng Wang, and Yun Shi. 2019. Cucumber leaf disease identification with global pooling dilated convolutional neural network. *Computers and Electronics in Agriculture* 162 (2019), 422–430.

[371] Wendong Zhang. 2021. The Case for Healthy US-China Agricultural Trade Relations despite Deglobalization Pressures. *Applied Economic Perspectives and Policy* 43, 1 (2021), 225–247.

[372] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. 2021. K-Net: Towards Unified Image Segmentation. In *NeurIPS*.

[373] Xiao Zhang and Ximing Cai. 2011. Climate change impacts on global agricultural land availability. *Environmental Research Letters* 6, 1 (2011), 014014.

[374] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2017. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv Preprint 1707.01083* (2017).

[375] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2021).

[376] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. 2018. M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network. *arXiv Preprint 1811.04533* (2018).

[377] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. 2021. Meta Label Correction for Noisy Label Learning. In *AAAI*, Vol. 35.

[378] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. 2019. Unequal-training for deep face recognition with long-tailed noisy data. In *CVPR*. 7812–7821.

[379] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. (2020), 1–8.

[380] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. 2021. DeepViT: Towards Deeper Vision Transformer. *arXiv preprint arXiv:2103.11886* (2021).

[381] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. 2019. Bottom-Up Object Detection by Grouping Extreme and Center Points. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[382] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review* 5, 1 (2018), 44–53.

[383] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16804–16815.